Resum

En aquest treball es dissenyen i desenvolupen dues eines, una a partir d'un programa i l'altra obtinguda a partir d'una macro d'Excel, amb l'objectiu d'agilitzar el procés de disseny d'amplicons i millorar l'anàlisi del rendiment dels mateixos al laboratori de genòmica del càncer del Vall d'Hebron Institut d'Oncologia.

Els amplicons són fragments d'ADN producte de l'amplificació produïda per un procés químic, que s'utilitzen per a facilitar la detecció de mutacions en mostres de pacients de càncer.

Prèviament al disseny i desenvolupament de les aplicacions, s'ha observat el procediment seguit en el laboratori per tal de trobar punts millorables i entendre la teoria que hi ha darrere del procés. A continuació s'han analitzat les alternatives existents a l'hora de desenvolupar les eines i s'han triat les opcions que s'han considerat més adients.

Seguidament s'ha procedit al disseny de l'estructura de les aplicacions i s'han desenvolupat en llenguatge de programació Visual Basic .NET. Amb aquestes aplicacions s'ha fet una prova amb el gen BRCA1 per verificar el seu funcionament. El resultat d'aquest experiment ha estat positiu i s'ha vist una millora important respecte el mètode anterior de disseny d'oligonucleòtids i el càlcul de la seva eficiència.

Com a resultat s'han desenvolupat dues eines que agilitzen el procés d'anàlisi de mostres de pacients al laboratori i s'ha millorat el mètode per a verificar l'eficiència dels amplicons implicats.

Sumari

1 Glossari

1.1 Abreviacions

ADN: àcid desoxiribonucleic

ARN: àcid ribonucleic

IDE: Entorn de desenvolupament integrat

IGV: *Integrative Genomics Viewer.*

Macro: macroinstrucció

NGS: *Next Generation Sequencing*

PCR: reacció en cadena de la polimerasa

SNP: polimorfisme d'un sol nucleòtid

1.2 Definicions

Donat que es tracta d'un treball multidisciplinar, s'ha considerat adient incloure un apartat de conceptes de biologia al final d'aquesta memòria.

La resta de conceptes, que sí són més propers al que s'ha estudiat al Grau són els que llisten a continuació.

Assay Design: programa desenvolupat per l'empresa americana Sequenom de cara a l'ús de la seva tecnologia per tal de detectar mutacions o SNPs a l'ADN genòmic. A aquest programa s'hi introdueixen seqüències genòmiques en format de text per tal que retorni els oligonucleòtids més adients per a poder amplificar les mencionades regions d'interès mitjançant la tecnologia PCR.

El fitxer d'entrada té una columna amb els noms de les regions d'interès i una columna on s'especifica la seqüència d'aquestes regions en lletres majúscules i minúscules. La Fig. 1.1 és un exemple de com apareix la seqüència en aquest fitxer. Les lletres majúscules serveixen per indicar al programa que aquelles són les bases en les qual es pot situar l'oligonucleòtid. En canvi, les lletres minúscules indiquen que aquella és la regió d'interès i que, per tant, l'oligonucleòtid no es pot situar allà.

TGGATTTATCTGCTCTTGCGCTTGAAGAAGTACAAaatgtcattaatgctatgcagaaaatcttagagtg[t/-]cccatctggtaagtcagcacaagagtgtattaatt**TGGGATTCCTATGATTATCTCCTATGCAAATGAAC**

Fig. 1.1 Exemple de seqüència que constaria com a entrada pel programa Assay Design. En negreta es poden veure les regions on es podran situar els oligonucleòtids, un al principi i un altre al final.

Bam: el format bam és la versió binària del format SAM.

Bed: és un format de text que proporciona flexibilitat per a definir línies de dades que es mostren en un conjunt. Cada línia pot tenir fins a 9 columnes que contenen informació. Tres d'aquestes columnes són obligatòries i les altres sis poden aportar informació addicional sempre i quan es mantingui el mateix format per a tots els camps del fitxer. En aquest treball, cada fila dels fitxers bed correspon a un amplicó o regió i inclouen, per ordre, el cromosoma, la coordenada inicial, la coordenada final (camps obligatoris) i el nom de la regió.

Disciplina imperativa: conjunt d'instruccions que indiquen al computador com realitzar una tasca en contraposició amb la disciplina declarativa, que requereix de l'especificació de condicions, equacions i transformacions entre d'altres elements, per a descriure un problema. La programació imperativa descriu pas a pas les instruccions que s'hauran d'executar per variar l'estat del programa i trobar la solució.

Entorn de desenvolupament integrat: aplicació que proporciona serveis integrals per facilitar el desenvolupament de software. Normalment consta d'un editor de codi de font, eines de construcció i un depurador.

Fals positiu: en general, és una resposta positiva a una prova o experiment que es deu a un error d'algun tipus perquè, en realitat, no hauria d'aparèixer. Al laboratori aquests casos es donen quan la polimerasa comet un error en la còpia d'una seqüència o per canvis químics produïts al llarg del temps mentre la mostra es guarda en forma de parafina. És molt important tenir present que aquests canvis es produeixen de forma aleatòria.

Fastq: format de text utilitzat per emmagatzemar seqüències biològiques (com seqüències d'oligonucleòtids) amb les seves puntuacions de qualitat corresponents. Tant la seqüència de lletres com el valor de qualitat estan codificades amb un sol caràcter segons el codi ASCII.

IGV: software de visualització per a l'exploració interactiva de grans bases de dades de genòmica. És compatible amb una àmplia varietat de tipus de dades, incloses les dades basades en NGS, i les anotacions del genoma. En aquest treball, s'utilitza per veure els gens en detall, fins arribar a la seqüència d'ADN i s'hi afegeixen fitxers bed per visualitzar els amplicons. Com que els camps d'aquest arxiu tenen el cromosoma i les coordenades de les regions, els fragments es representen a la seva posició real en el genoma.

MiSeq: seqüenciador d'ADN de la companyia Illumina que s'utilitza al laboratori per a generar seqüències de les regions d'interès i posteriorment detectar-ne les mutacions. Com a seqüenciador, té la funció de determinar l'ordre de nucleòtids de les seqüències, és a dir, de donar com a sortida la seqüència corresponent a una mostra d'entrada.

Pipeline: en informàtica, és un conjunt d'elements de processament de dades connectats en sèrie de forma que la sortida de cada fase és l'entrada de la següent. Aquesta arquitectura és molt comú en el desenvolupament de programes per a la interpretació de comandes, donat que es poden concatenar ordres fàcilment mitjançant "canonades" (pipe).

Programació funcional: tipus de programació que es basa en l'ús de funcions matemàtiques, al contrari que en el cas de la programació imperativa.

Programació orientada a objectes: tipus de programació que utilitza objectes (entitats amb una sèrie d'atributs de funcionalitat i comportament) a les seves interaccions.

PuTTY: és un client de Shell segur de llicència lliure que s'utilitza al laboratori per a connectar-se al servidor Linux del laboratori des d'un equip Windows i així llençar les pipelines des de la línia de comandament.

Sam: el format SAM (Sequence Alignment Map) és un format de text delimitat per tabulacions que s'utilitza per emmagatzemar grans quantitats de seqüències d'oligonucleòtids. Proporciona la flexibilitat necessària per poder emmagatzemar tota la informació d'alineació de seqüències i alhora és prou senzill per permetre que la seva generació sigui simple per part dels programes d'alineació.

Script: els scripts són programes, generalment simples, que s'utilitzen per a realitzar tasques molt específiques. Normalment aquestes instruccions s'emmagatzemen en arxius de text que s'hauran d'interpretar línia per línia en temps real. Per aquests usos és freqüent que els shells (intèrprets d'ordres) siguin els intèrprets d'aquest tipus de programes.

Seqüenciador: màquina que té com a objectiu determinar l'ordre dels nucleòtids en un tram d'ADN.

2 Prefaci

2.1 Origen del projecte

En les últimes dècades l'estudi molecular del càncer ha esdevingut un camp important en la investigació mèdica, donant cada cop més importància a la comprensió d'aquesta malaltia que es preveu que augmenti en un 14% la seva incidència com a mitja a la Unió Europea i un 18% a l'Estat espanyol [21].

El càncer és una malaltia genòmica, és a dir, el seu origen es troba en alteracions o mutacions del genoma. De forma simplificada, el desenvolupament d'un tumor esdevé per l'acumulació i selecció de certes mutacions (canvis estables)al genoma del teixit sa originari. Algunes d'aquestes mutacions es seleccionen recurrentment en els tumors i són conegudes com a "drivers" i contribueixen activament en el desenvolupament del càncer. De fet, en alguns tipus tumorals, s'han identificat mutacions que es presenten freqüentment i que es poden inhibir amb un tractament farmacològic (alguns reconeguts per entitats oficials i altres en procés de validació). És en aquest escenari on es fa imprescindible desenvolupar tècniques que permetin identificar aquests conjunts de mutacions amb possibilitat de tractament pels pacients.

Les noves tecnologies de seqüenciació massiva (NGS) han permès la obtenció d'una quantitat ingent de seqüències per experiment (fins a 600 milions) i, per tant, han esdevingut una eina emprada en els últims anys per a la caracterització del genoma del càncer a nivell poblacional i, en els darrers temps, a nivell del pacient individual.

El laboratori de Genòmica del Càncer del Vall d'Hebron Institut d'Oncologia ha desenvolupat una aplicació per tal de fer servir les tecnologies NGS en la detecció de mutacions en 61 gens, basant-se en la seqüenciació de més de 800 regions en aquests

gens, que són amplificades mitjançant la tècnica PCR de forma paral·lelitzada (PCR multiplexada). Aquest tipus d'aplicacions basades en l'obtenció i seqüenciació de productes de PCR es coneixen de manera general com "Amplicon-Seq".

El desenvolupament de nous panells de PCR multiplexada que permetin la seqüenciació de nous gens és un camp en continu avanç al laboratori, a mesura que es genera la necessitat des de l'àmbit clínic, amb nous fàrmacs en desenvolupament. Una part d'aquesta memòria descriurà aquest procés i la seva optimització.

L'anàlisi posterior a la seqüenciació permet conèixer les mutacions que presenten les mostres de tumors i els resultats ajuden als oncòlegs tractar adequadament i, per tant, millorar la qualitat i expectativa de vida dels pacients.

Un cop establerta la importància de la detecció de mutacions en el tractament del càncer i després de formar part del laboratori de genòmica del càncer de la Vall d'Hebron durant més d'un any, sorgeix l'oportunitat de millorar alguns aspectes i aprofundir en els coneixements d'aquest camp.

A partir d'aquesta voluntat es decideix observar tot el procés d'una de les tècniques més comuns al laboratori, l'Amplicon-Seq, per tal de trobar punts febles en els quals es pugui aportar alguna millora.

2.2 Motivació

A l'interès en adquirir nous coneixements i millorar el rendiment de la pròpia feina com a motivacions principals d'aquest treball, s'hi afegeix la gran importància d'una malaltia com el càncer en l'àmbit de la salut, així com un al·licient especial de caire personal.

Comprendre les bases de la feina que es desenvolupa al laboratori per a poder analitzar el procediment i trobar-ne punts millorables suposa un repte. S'ha observat que el procediment actual de disseny d'oligonucleòtids (seqüències d'ADN que es descriuen al glossari d'aquesta memòria) és lent i no permet satisfer la demanda creixent de nous panells d'amplicons. A més a més, s'ha advertit un problema a l'hora d'identificar quins d'aquests amplicons tenen mal funcionament. És per aquest motiu que sorgeix la necessitat de desenvolupar eines que ajudin a millorar el rendiment de la feina i permetin realitzar les tasques en menys temps.

A més a més, en una societat cada cop més vinculada a eines informàtiques per a treballar, resulta important per a una persona de l'àmbit tècnic assolir coneixements per tal de poder generar petits programes o recursos que ajudin en el funcionament d'una empresa o una organització. Per aquest motiu serà enriquidor aprendre nous llenguatges de programació.

Ser capaç de realitzar un projecte des de la concepció de la idea inicial fins a trobar una solució i extreure'n unes conclusions és un desafiament, així com aprendre a barrejar conceptes assolits al Grau d'Enginyeria en Tecnologies Industrials juntament amb nous coneixements per tal d'arribar a assolir els objectius plantejats.

Per últim, suposa una satisfacció personal el fet de veure els progressos propis en un nou entorn de treball en el que ha estat necessari adquirir nous coneixements.

2.3 Requeriments previs

Per a realitzar aquest treball ha estat necessari disposar de coneixements com programació informàtica, química, gestió de projectes, ús avançat de Microsoft Excel i anglès per a poder accedir a informació útil i d'actualitat.

També han estat d'utilitat altres conceptes assolits al Grau en Enginyeria de Tecnologies Industrials, com l'ús solvent dels recursos d'informació o la capacitat de desenvolupar-se en un àmbit nou i trobar solucions als inconvenients que sorgeixen.

Per altra banda, ha estat necessari assolir coneixements bàsics de genòmica per tal de comprendre el funcionament dels amplicons i entendre la lògica que hi ha darrere del procés que es realitza al laboratori.

3 Introducció

3.1 Objectius del projecte

En aquest projecte s'han diferenciat dos grups d'objectius:

- Objectius principals: en aquesta part es té en compte l'objecte inicial del projecte, solucionar els problemes observats (la necessitat que el disseny d'amplicons sigui més ràpid i de millorar el càlcul de cobertura per veure l'eficiència d'aquests amplicons).
- Objectius específics i funcionals: són aquells que responen a les necessitats dels usuaris.

3.1.1 Objectiu principal

En aquest treball l'objectiu principal és el de desenvolupar eines per tal d'agilitzar el procés de disseny i calcular l'eficiència dels amplicons de forma més acurada.

Cal esmentar que en el moment de realització d'aquest projecte no s'ha trobat cap eina disponible per a realitzar la fragmentació de seqüències, que és el pas més pesat i lent a l'hora de dissenyar amplicons.

Pel que fa al càlcul de l'eficiència, s'ha observat que el mètode que s'utilitzava fins ara dóna resultats poc acurats i provoca confusions. Per aquest motiu forma part dels objectius principals trobar una forma de millorar-ho.

També forma part dels objectius inicials el fet d'assolir nous coneixements de l'àmbit de la informàtica, com seria un nou llenguatge de programació i aprendre el procediment a seguir davant d'una situació inicialment desconeguda per tal de trobar solucions als problemes que sorgeixen.

3.1.2 Objectius específics i funcionals

Es defineixen com a usuaris els tècnics del laboratori que utilitzaran les diverses aplicacions desenvolupades. A partir d'aquesta premissa es determinen els següents objectius:

- Les eines han de tenir un ús senzill i apte per a persones sense coneixements de programació.
- Portabilitat: la instal·lació ha de ser senzilla per a tots els equips del laboratori.

- Els paràmetres de disseny d'amplicons han de poder canviar fàcilment en el cas que es vulgui aplicar a màquines de requeriments diferents.
- La interfície de l'aplicació ha d'adequar-se al context en el que s'emmarca i ha de tenir un disseny clar que permeti als usuaris un ús intuïtiu i ràpid.
- Les eines que es desenvolupin no poden ocupar gaire espai de memòria ni requerir d'un equip potent donat que cada usuari del laboratori té un ordinador de diferents característiques.

3.2 Abast del projecte

Aquest treball inclou tot el procés de concepció, disseny i programació de les eines descrites anteriorment així com la prèvia documentació i assoliment dels coneixements biològics implicats en la memòria.

També s'inclou dins de l'abast del projecte l'aplicació de les eines dissenyades en un cas pràctic per a validar el seu funcionament i la seva utilitat. Amb l'ajut del fragmentador de seqüències desenvolupat, es dissenyaran els amplicons necessaris per a poder amplificar les regions d'un gen determinat i s'analitzaran els resultats per a poder aplicar l'eina que ajudarà a avaluar l'eficiència de cada amplicó.

No s'hi inclourà, però, la incorporació d'aquestes eines a cap portal obert a altres usuaris ni a mercats d'aplicacions donat que les eines s'han dissenyat per a satisfer les necessitats específiques d'un laboratori en concret. De totes maneres, en un futur es podria considerar la possibilitat d'ampliar les opcions dels programes per adequar-se a necessitats més generals i millorar la interfície amb l'objectiu de facilitar recursos d'ús lliure a personal d'investigació.

4 Conceptes previs i metodologia

4.1 El càncer

En aquest apartat es farà una breu introducció bàsica a la biologia que hi ha darrere del càncer per tal de poder entendre millor els conceptes i processos que apareixeran al llarg d'aquesta memòria.

La paraula càncer engloba un conjunt de malalties genètiques en les quals s'observa una divisió descontrolada de les cèl·lules del cos que adquireixen capacitat invasiva a altres localitzacions de l'organisme. Així doncs, primer es genera el tumor o neoplàsia, que prolifera i genera metàstasis.

Hi ha un conjunt de processos que s'han de descontrolar a nivell cel·lular per tal d'arribar al desenvolupament d'un tumor, coneguts com *Hallmarks of Cancer* (Característiques del càncer). Aquests processos són:

- Independència dels senyals de creixement: els senyals de creixement són la forma de comunicació que es dóna en teixits normals i que proporcionen la senyal de divisió de les cèl·lules. Per exemple, quan s'ha produït dany en un teixit, és necessari que les cèl·lules del voltant es reprodueixin per reomplir l'espai de les cèl·lules mortes.
- Insensibilitat a senyals en contra del creixement: existeixen senyals que es donen en els teixits normals i que donen senyal a les cèl·lules per no dividir-se i multiplicar-se. En l'exemple anterior, un cop reomplert l'espai de les cèl·lules mortes, cal aturar la reproducció.
- Evasió de l'apoptosi: el mecanisme conegut com apoptosi s'encarrega d'activar un programa de suïcidi cel·lular. Aquest mecanisme és freqüent, per exemple, en el desenvolupament embrionari.
- Potencial replicatiu infinit: les cèl·lules, al llarg de divisions successives, van perdent els extrems dels cromosomes fins que s'arriba a un punt on la cèl·lula ja no pot dividir-se més. Des del punt de vist biològic, aquest fet assegura que les cèl·lules massa velles i que han acumulat més danys no es propaguin en l'organisme.
- Angiogènesi sostinguda: l'angiogènesi és la creació de vasos sanguinis, és a dir, la possibilitat d'obtenir nutrients i oxigen per a les cèl·lules. En els tumor, s'activa l'angiogènesi per poder sostenir la divisió cel·lular descontrolada (que requereix de molts recursos energètics).

- Invasió tissular i metàstasi: els tumors són capaços d'ocupar espais més enllà dels límits naturals del teixit i, posteriorment, envair altres òrgans o estructures del cos.

Aquests canvis en el comportament de la cèl·lula tenen origen en un conjunt de mutacions genètiques que es poden produir, principalment, en tres tipus de gens segons la seva funció: proto-oncogens i gens supressors tumorals.

- Proto-oncogens: promouen el creixement i divisió de les cèl·lules. En alguns casos l'alteració d'aquest tipus de gens permet que les cèl·lules creixin i sobrevisquin més del que haurien.
- Supressors tumorals: s'encarreguen d'impedir la reproducció cel·lular excessiva. La mutació en aquests gens fa que perdin la seva funció i, per tant, que augmentin les probabilitats de desenvolupar un tumor.

A mesura que ha augmentat el coneixement científic sobre les mutacions i el canvis en les cèl·lules que produeixen, s'ha aconseguit que a dia d'avui l'enteniment dels tumors no sigui només en base a la seva localització, sinó a les alteracions genètiques que mostren.

4.1.1 De l'ADN al tumor

L'ADN, químicament, és una doble cadena antiparal·lela d'àcid desoxiribonucleic. Entre les seves funcions destaquen l'emmagatzematge d'informació (gens), la codificació de proteïnes (transcripció i traducció) i la replicació.

La codificació de proteïnes és una de les funcions més rellevants de l'ADN, donat que aquestes són les efectores funcionals a nivell cel·lular (les proteïnes són responsables de la majoria de funcions a la cèl·lula). De forma simplificada, el procés comença amb la transcripció de l'ADN. La transcripció es produeix al nucli de la cèl·lula i és el moment en el qual se sintetitza ARN utilitzant la seqüència d'ADN com a motlle. En primer lloc la cadena doble d'ADN es separa, permetent que l'enzim ARN polimerasa s'adhereixi a la cadena simple, tal i com s'observa a la Fig 4.1.

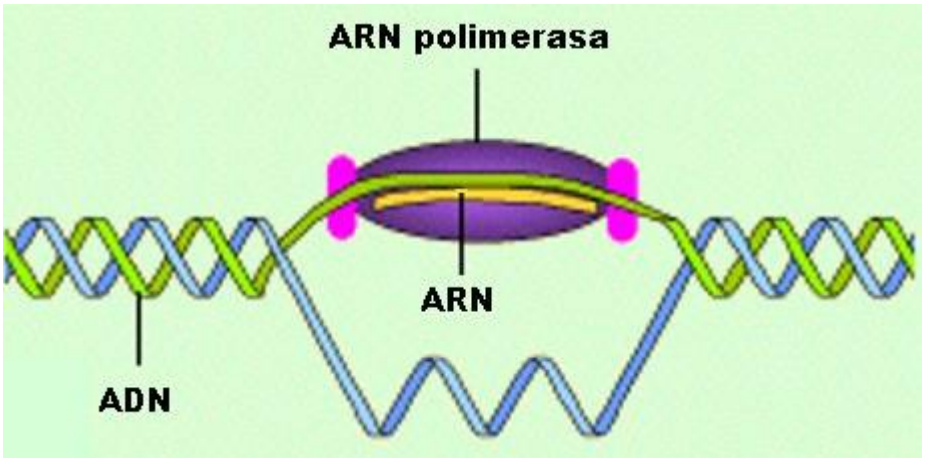

Fig. 4.1 Esquema de l'adherència de la polimerasa a la cadena simple d'ADN

En presència de nucleòtids lliures, l'ARN polimerasa s'encarrega de copiar la seqüència d'ADN per generar ARN missatger. Cal recordar que la còpia serà complementària (A amb U i C amb G).

L'ARNm generat encarà no està llest per a sortir al citosol, requereix d'un procés que s'anomena maduració. A l'inici de la maduració es realitza el que s'anomena splicing, que és un procés en el que es separen els introns i els exons de la seqüència i s'uneixen únicament els segments que contenen informació per formar proteïnes (exons). D'aquesta forma s'obtenen tots seguits.

Després de l'splicing és necessari que una seqüència especial de nucleòtids s'uneixi a un extrem de l'ARN. Aquesta estructura s'anomena caputxó, conegut com CAP en anglès i manté l'estabilitat en la traducció de l'ARN.

A l'altre extrem de l'ARN s'afegeix un grup anomenat poli A (poliadenina), un conjunt de bases A que també intervé en l'estabilitat per a sobreviure al citosol.

Després, arriba el procés de traducció que es produeix als ribosomes, continguts al citosol de la cèl·lula. En aquest pas l'ARNm s'adhereix al ribosoma i, amb l'ajuda de l'ARNt comença a generar aminoàcids a partir de grups de 3 bases, que s'anomenen codons. Un enzim s'encarrega d'anar unint els aminoàcids que s'obtenen, generant una cadena lineal. A la Fig. 4.2 es pot veure una taula de correspondència entre els triplets de bases i l'aminoàcid corresponent.

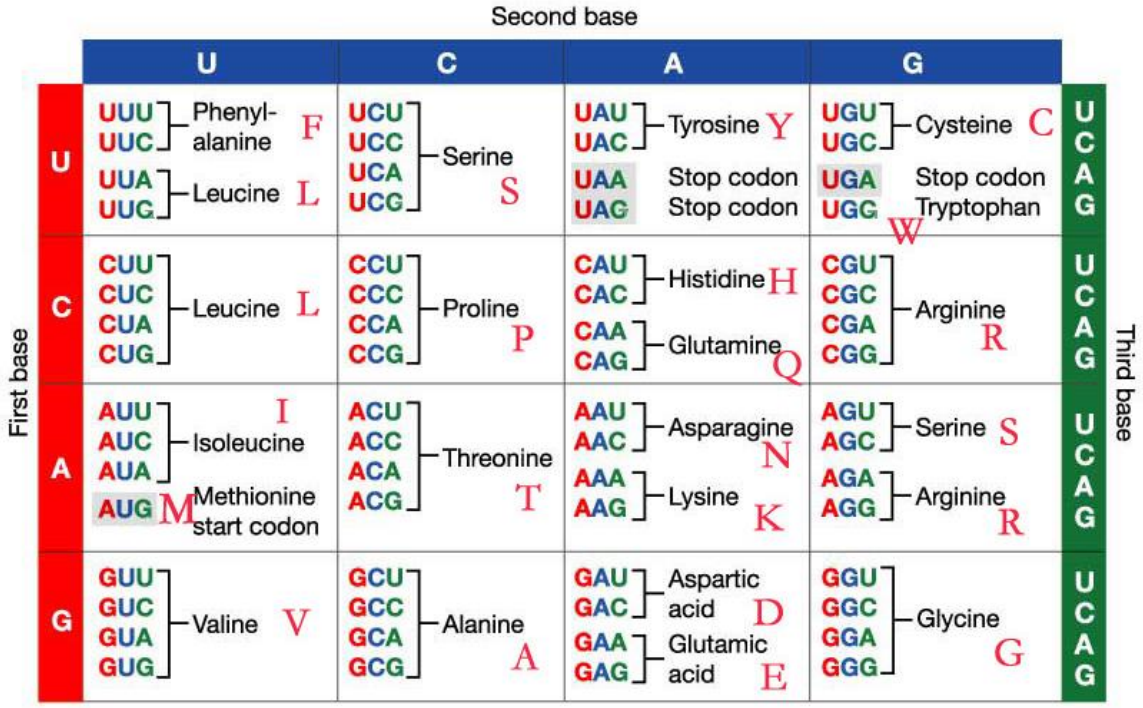

Fig. 4.2 Taula d'equivalència dels codons amb els aminoàcids corresponents

Finalment, aquestes cadenes donen lloc a les proteïnes, cadascuna amb una composició i longitud diferents. Les proteïnes realitzen múltiples funcions entre les quals destaquen:

- Funció estructural: formen l'arquitectura de la cèl·lula.
- Funció immunològica: defensa contra agents externs, un exemple són els anticossos.
- Enzims: són proteïnes amb activitat biològica que permeten accelerar reaccions químiques
- Proteïnes contràctils: són les responsables de les contraccions dels músculs
- Protecció (no immunològica): actuen com a barrera. Per exemple, quan hi ha una ferida es crea una pel·lícula per interrompre el sagnat.

4.1.2 Importància de la genòmica en l'estudi del càncer

Com s'ha vist, l'ADN juga un paper essencial en la síntesi de proteïnes. Una variació en la seqüència implica canvis en la seqüència d'ARNm que s'obtindrà després de la transcripció i, per tant, influirà en les proteïnes que es sintetitzaran a les cèl·lules.

D'aquesta manera, una mutació pot implicar canvis en les proteïnes sintetitzades i, canviar les propietats de la cèl·lula. Aquests canvis són els que donen lloc als tumors.

4.2 La PCR

El procés de PCR (*polymerase chain reaction)* és una tècnica que va ser desenvolupada per Kary B. Mullis el 1989, atorgant-li el premi Nobel de Química el 1993.

Aquesta tècnica permet obtenir un gran nombre de còpies idèntiques d'un fragment d'ADN. Per tal de direccionar quina regió del genoma es vol copiar, s'utilitzen cebadors específics per a la síntesi d'ADN, catalitzada per un enzim anomenat Taq Polimerasa.

El procés de la PCR consisteix en una sèrie de cicles que inclouen 3 fases esquematitzades a la Fig. 4.3:

- Desnaturalització: per tal que la reacció es pugui iniciar, és necessari que la cadena d'ADN es presenti en forma de cadena simple. En aquesta fase es produeix una ruptura dels ponts d'hidrogen que hi ha entre les cadenes mitjançant un increment de la temperatura fins als 90-95ªC i mantenint-la durant uns minuts.
- Hibridació: també denominada *annealing*, és una fase en la qual la temperatura disminueix fins als 40-60ºC per tal que els oligonucleòtids s'uneixin a les regions d'interès. La temperatura adient dependrà de diversos factors i és relativament específica per a cada oligonucleòtid. Un dels factors més determinants serà la longitud dels oligonucleòtids, així com la seva seqüència.
  Una forma simplificada per a calcular aquesta temperatura (temperatura de fusió) és:
  $Tm = 4(G + C) + 2(A + T)$, on *G, C, A, T* corresponen al nombre d'aquestes bases presents en dita seqüencia.
- Extensió: en aquest pas la polimerasa afegeix nucleòtids (A, T, C, G) a un dels extrems de l'oligonucleòtid seguint com a guia la cadena d'ADN desnaturalitzada.

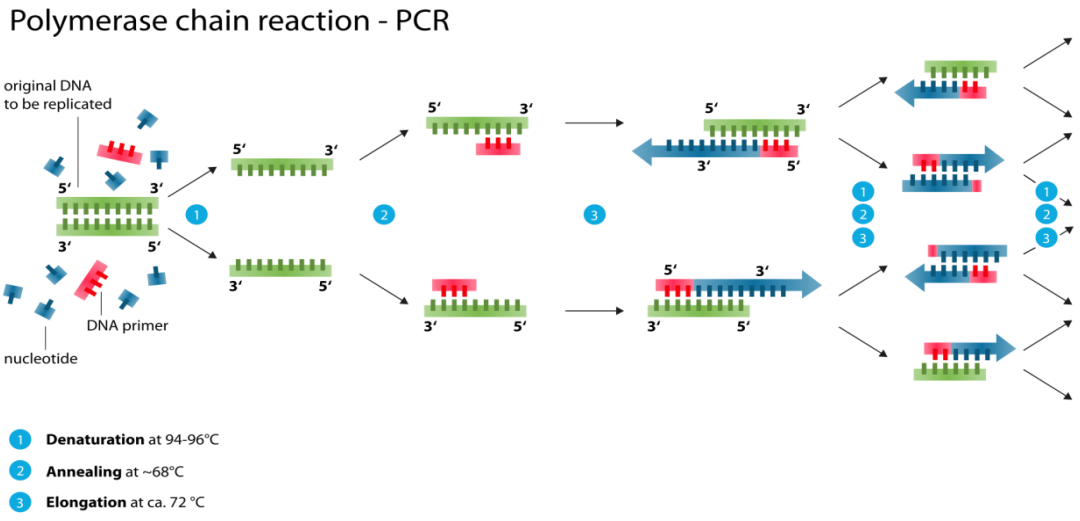

Fig. 4.3 Esquema de funcionament de la PCR

4.2.1 PCR Multiplexada

Dins del camp de les reaccions en cadena de la polimerasa podem trobar una sèrie d'aplicacions que s'han desenvolupat i, entre elles, la PCR multiplexada.

Aquesta modalitat de PCR permet diverses amplificacions simultànies de diferents seqüències d'interès en una única reacció química. Això s'aconsegueix utilitzant diverses parelles d'oligonucleòtids que funcionen en paral·lel.

Aquesta tècnica accelera el procés d'amplificació de la mostra d'un pacient però requereix d'una sèrie d'aspectes a tenir en compte:

- Longitud dels oligonucleòtids: aquest tipus d'assaig de PCR comporta el disseny d'un gran nombre d'oligonucleòtids. Per tant, és important que aquests tinguin una longitud semblant per tal que les temperatures de fusió puguin ser uniformes.
  La longitud també afecta a altres aspectes importants en una reacció d'aquest tipus. Per exemple, uns oligonucleòtids massa curts implicaran poca especificitat en l'amplificació.
- Especificitat: és important considerar la possibilitat de que els oligonucleòtids s'adhereixin a regions diferents a les que es prenien com a objectiu. Aquest fet obliga a dissenyar seqüències d'oligonucleòtids que disminueixin al màxim el risc de que es doni aquesta situació i és per això que aquestes compten amb una combinació de bases prou llarga i única.
- Temperatura d'hibridació: és important que tots els oligonucleòtids d'un mateix well presentin temperatures d'hibridació semblants per tal d'evitar que les diferències generin problemes en el procés d'hibridació.
  Si un d'ells presenta una temperatura força més baixa que la resta dels que hi ha a la mateixa reacció, la unió es farà de forma poc específica mentre que, en el cas contrari, no es produiria una unió completa.
- Evitar la formació de dímers: aquest concepte es refereix a la possibilitat de que el oligonucleòtids dissenyats puguin hibridar-se entre ells pel fet que les bases que els formen siguin complementàries. Aquesta situació provocaria una amplificació dels oligonucleòtids mateixos i implicaria problemes en la posterior quantificació de lectures de la PCR.

4.3 Funcionament del MiSeq de l'empresa Illumina. Synthesis-by-sequencing

El funcionament del seqüenciador emprat al laboratori es basa en la Synthesis-by-sequencing (seqüenciació per síntesi) que consisteix en el procés que es descriu a continuació.

En aquest procés, els fragments esmentats anteriorment (juntament amb els adaptadors) s'uneixen a la superfície d'una petita placa que s'introdueix a la màquina. Com es pot veure a la Fig. 4.4, aquesta superfície disposa d'una sèrie de fragments que són complementaris als adaptadors que s'han inclòs a les llibreries.

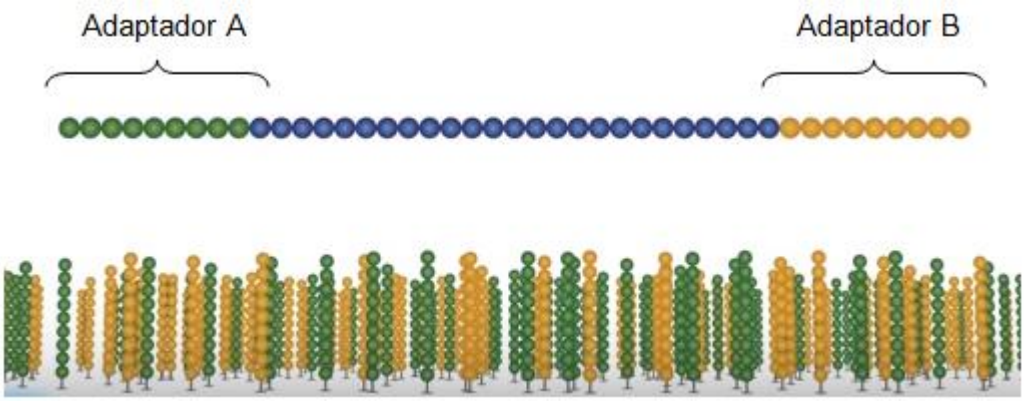

Fig. 4.4 Primera imatge del procés intern al seqüenciador. Es pot veure que els adaptadors es corresponen amb petites seqüències adherides a la superficie de la placa.

La polimerasa genera una còpia complementària del fragment de forma que l'original es pot eliminar. En el següent pas el bri d'ADN es doblega de forma que l'adaptador B s'uneix amb el seu complementari tal i com es pot veure a la Fig. 4.5. De nou es genera una còpia de la seqüència.

Finalment es desnaturalitza i s'obtenen dues còpies complementàries del fragment original adherides a la placa per separat. Aquest procés es repeteix successivament de forma que s'obtenen milions de còpies. Finalment només es mantindran aquelles amb la seqüència idèntica a l'original, és a dir, les complementàries s'eliminaran.

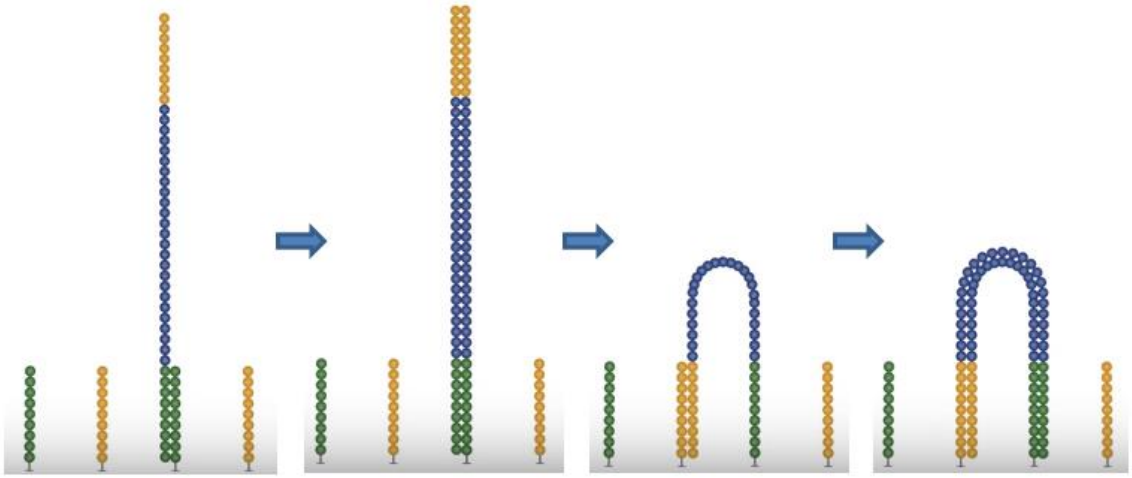

Fig. 4.5 Esquema de procés de duplicació de la seqüència d'interès

Un cop s'han obtingut totes les còpies, es procedeix a identificar-ne la seqüència. Com que l'ordre de bases de l'adaptador és coneguda, es començarà adherint el complementari d'aquest fragment de seqüència per tal d'assegurar que la posterior reacció comenci i identificar que s'inicia una lectura (primer pas de la Fig. 4.6). A partir d'aquí un corrent d'oligonucleòtids lliures competirà per adherir-se a continuació de la seqüència però només un d'ells serà l'adient (basant-se en la complementarietat amb la seqüència original).

Els nucleòtids emprats tenen l'especial característica d'emetre llum en el moment d'unió, de forma que emetran un color diferent segons si són A, C, T o G. D'aquesta forma només caldrà que la màquina identifiqui el color emès en cada instant per conèixer la seqüència. Aquest procés s'anomena seqüenciació per síntesi.

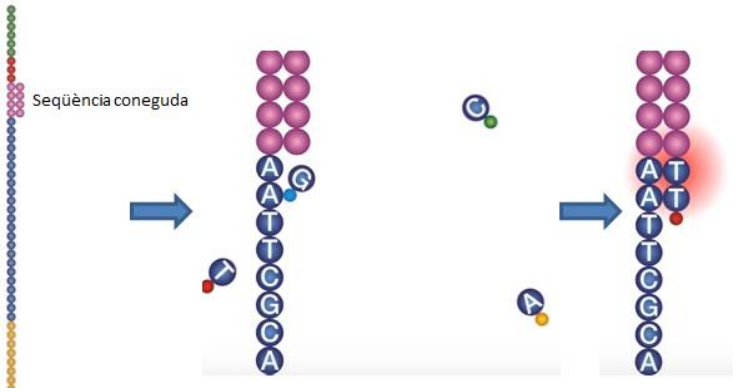

Fig. 4.6 Esquema d'unió d'oligonucleòtids lliures amb la respectiva emissió lumínica

Gràcies a l'amplificació prèvia que es realitza, s'emetran molts senyals de llum alhora, de forma que els sensors de la màquina podran detectar l'emissió amb més facilitat. L'emissió de llum d'un sol nucleòtid no seria detectable pels sensors de la màquina.

Donat que a l'extrem final de la seqüència hi ha l'altre adaptador conegut es pot conèixer on està el final de la lectura. En aquest moment, es treu el producte que s'ha obtingut adherint nucleòtids lliures.

La seqüència que fa de motlle es plega tal i com es pot veure a la Fig. 4.7 i es copia de nou de forma complementària. Les dues seqüències es separen i l'original es retira.

Es repeteix de nou tot el procés de lectura i finalitza el pas d'addició de nucleòtids. D'aquesta manera s'han obtingut dues lectures complementàries a partir de la seqüència inicial.

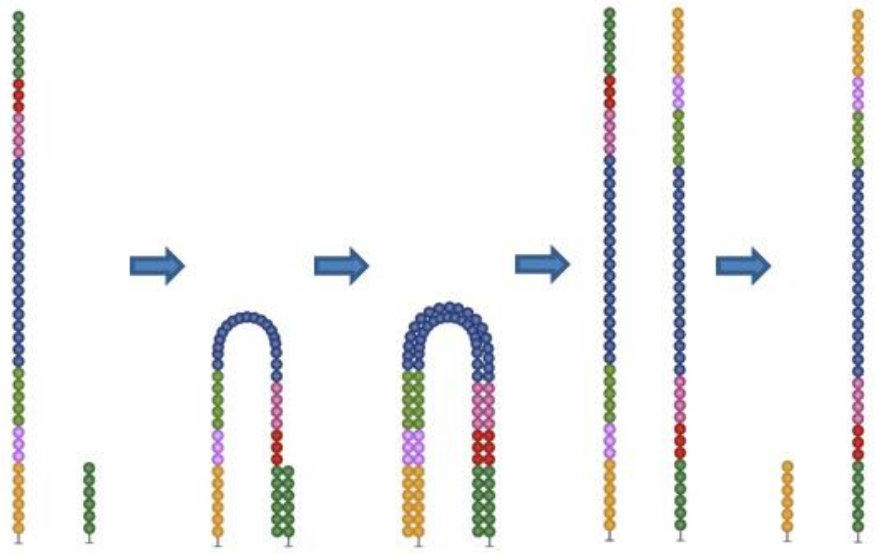

Fig. 4.7 Esquema de d'obtenció dela seqüència complementària

En aquest punt, totes les lectures s'ajunten i es mostren en un fitxer de format fastq.

5 Procediment actual al laboratori

Aquest treball es desenvolupa al laboratori de Genòmica del Gàncer del Vall d'Hebron Institut d'Oncologiaon actualment es realitzen múltiples tasques, entre les quals destaca l'anàlisi de mostres de pacients amb amb tècniques d'última generació. Les mostres s'analitzen amb l'objectiu d'identificar-ne mutacions que donin informació sobre tractaments o puguin permetre la inclusió dels pacients en assajos clínics amb fàrmacs en desenvolupament.

Esquemàticament els passos són:

- Disseny d'un panell d'Amplicon-Seq: es dissenya un conjunt d'oligonucleòtids per a realitzar la PCR multiplexada en regions d'interès i posteriorment es valida a nivell de laboratori per tal de poder-ho utilitzar en pacients. Aquest procés és recursiu, donat que poden aparèixer noves regions d'interès i caldrà ampliar el panell o dissenyar-ne un de nou.
- Establiment d'una rutina on s'aplica l'eina de PCR multiplexada en les mostres de pacients.

Per tal de facilitar la comprensió del projecte, es detallarà en primer lloc la rutina al laboratori i, posteriorment, la fase de disseny d'oligonucleòtids per a la PCR multiplexada, que és un dels àmbits de millora d'aquest treball.

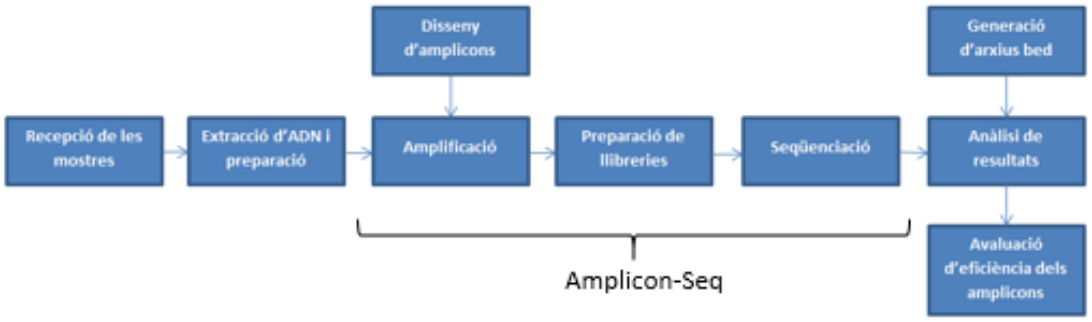

Fig. 5.1 Diagrama de blocs del procés que es segueix des de la recepció de les mostres fins a l'anàlisi i l'avaluació d'eficiència d'amplicons

5.1 Rutina d'anàlisi de mutacions en pacients mitjançant Amplicon-Seq

El procediment que es segueix actualment al laboratori comença per la recepció de les mostres juntament amb la corresponent sol·licitud d'anàlisi. El personal tècnic comprova que les dades del formulari siguin correctes i procedeixen a l'extracció d'ADN de les mostres.

Un cop extret i després d'una preparació prèvia, l'ADN es sotmet a un procés d'amplificació que es porta a terme mitjançant la PCR multiplexada amb amplicons específics per a regions d'interès. D'aquesta forma s'obté suficient material copiat del tumor original per poder-lo seqüenciar. Cal tenir en compte que la seqüenciació es fa amb una màquina que capta senyals lumínics. Si no hi hagués prou còpies, l'emissió de llum seria indetectable.

Amb els productes obtinguts, es preparen les llibreries (productes d'un procés que consisteix en lligar, a ambdós extrems dels fragments, uns adaptadors que seran necessaris al procés de seqüenciació).

El següent pas és la seqüenciació de l'ADN de les mostres amb el seqüenciador MiSeq de l'empresa Illumina, que s'ha descrit anteriorment.

A partir de l'obtenció dels arxius fastq amb les seqüències, comença la tasca dels bioinformàtics, que realitzaran l'anàlisi pertinent per tal d'extreure les mutacions presents a les mostres tumorals. L'anàlisi comença amb la preparació dels fitxers necessaris per poder llençar la pipeline, on es realitzen una sèrie de càlculs que condueixen al resultat final. L'esquema de la Fig 5.2 reflexa el camí que segueixen les dades des del moment en que surten de la màquina fins que s'obté l'informe de resultats.

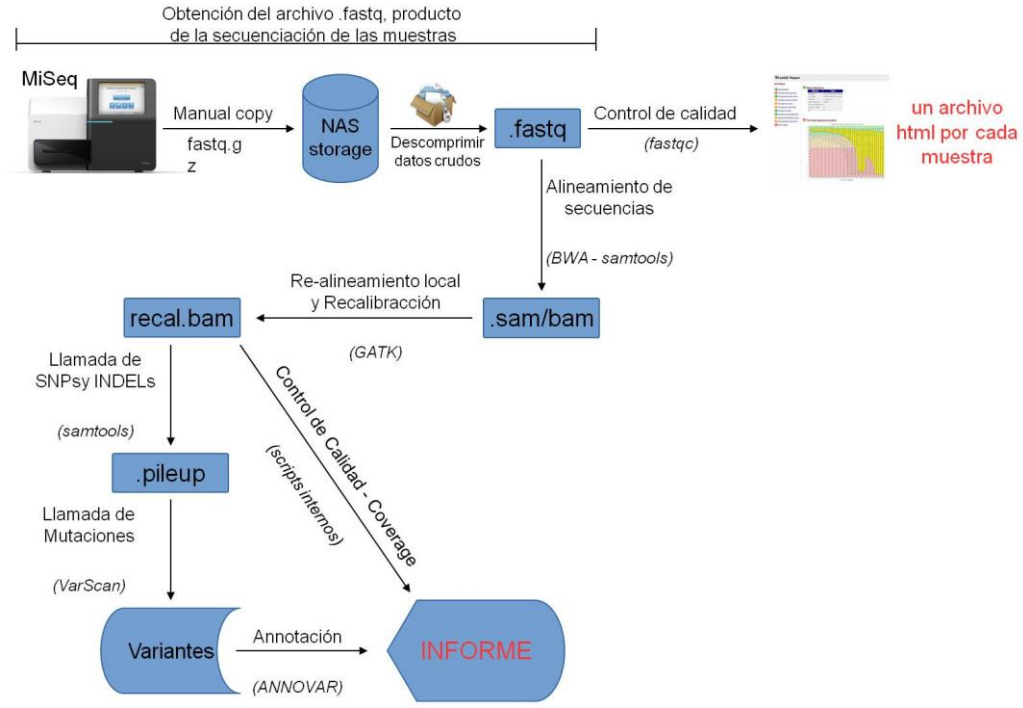

Fig. 5.2 Diagrama de blocs del flux de les dades

En primer lloc es copien les dades obtingudes de la màquina en format fastq a una de les unitats d'emmagatzematge del laboratori.

Un cop copiades, es crea una carpeta nova corresponent a aquest experiment i es copien els scripts Shell de la pipeline. En aquesta carpeta també s'hi inclouran tres fitxers de text que faran referència a la correspondència entre els números de mostra i els pacients a qui corresponen, així com la localització de les dades.

Després de realitzar aquesta tasca, es procedeix a llençar la pipeline, anomenada *pipeline_ampliconseq.sh*. Com a script inicial de la pipeline, té la funció de cridar els següents scripts, gestionar la descompressió de les dades d'entrada (fastq.gz) i copiar-les a la carpeta de sortida.

El següent script és l'anomenat *bucle_vc_clia.sh* i té la funció de fer el control de qualitat, eliminar les lectures de baixa qualitat, eliminar els adaptadors que s'han adherit als oligonucleòtids en passos anteriors, la eliminació de lectures massa curtes, l'alineament contra el genoma humà (obtenint els fitxers .sam/bam) i la creació dels fitxers recal.bam i pileup. Finalment crida el següent script *pileup2cns_amp.sh*

En aquest script es criden les variants, es filtren segons una base de dades de variacions genètiques humanes i s'anoten.

El següent script, *demultiplex_sinopsis.sh,* filtra els falsos positius recurrents i crea l'informe de variants, que rep el nom de "sinopsis_varscan_filtered.xls".

L'script *Coverage.sh* calcula una sèrie de factors de qualitat per cada mostra (mitjana de cobertura, cobertura total, nombre de lectures abans de filtrar, percentatge de regions cobertes correctament) i crea un informe anomenat "AmpCoverage.txt". Finalment, l'script *CoverageDuplicates.sh* calcula els mateixos factors de qualitat per a cada pacient (cal esmentar que a cada pacient li corresponen dues mostres).

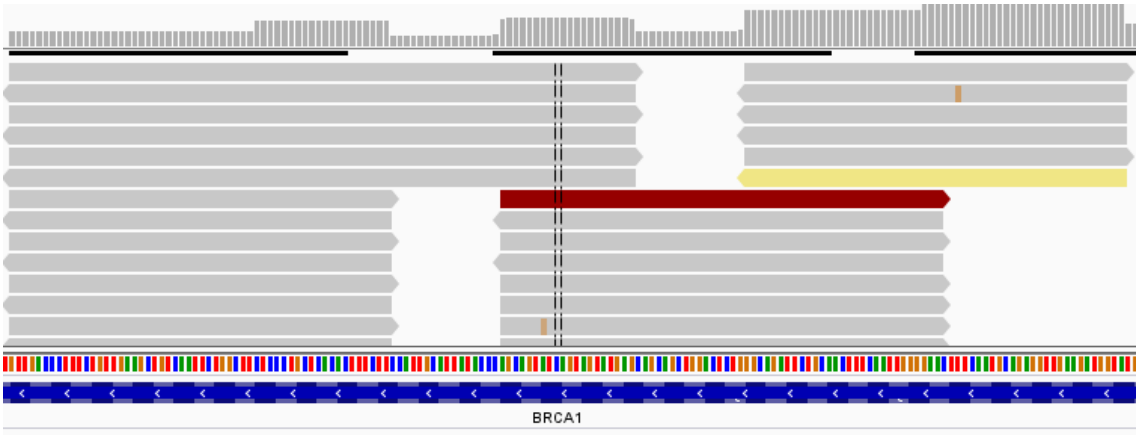

Fig. 5.3 Imatge de l'IGV on es veu a la part superior un gràfic de barres corresponent al nombre de lectures detectades a cada punt. A la resta de la imatge es poden veure les lectures que ha realitzat la màquina

Com s'observa a la Fig 5.4, al final del procés de la pipeline s'obtenen tots els fitxers necessaris per analitzar els resultats i el funcionament dels amplicons.

Les possibles mutacions detectades s'analitzen amb l'ajuda del software IGV, que permet visualitzar de forma esquemàtica la distribució i magnitud dels senyals obtinguts al fitxer recal.bam (tal i com es pot observar a la Fig. 5.3). A partir d'una sèrie de paràmetres establerts es determina si les mutacions són vàlides o si, pel contrari, es tracta de falsos positius. També es té en compte el fet que, de cada mostra d'un pacient, se'n fan duplicats. D'aquesta manera es pot veure quines alteracions de la freqüència s'han produït per canvis químics aleatoris (hi ha molt poca probabilitat que apareguin en els dos duplicats) i descartar-les com a falsos positius.

Els resultats de les anàlisis s'agrupen en un fitxer que, després de ser supervisat per la cap del laboratori, arriba als oncòlegs o investigadors per tal de conèixer les mutacions detectades a cada pacient.

Les dades de cobertura obtingudes en el fitxer “AmpCoverage.txt” s’utilitzen per veure quines regions no funcionen correctament i estudiar si és necessari un redisseny dels amplicons implicats.

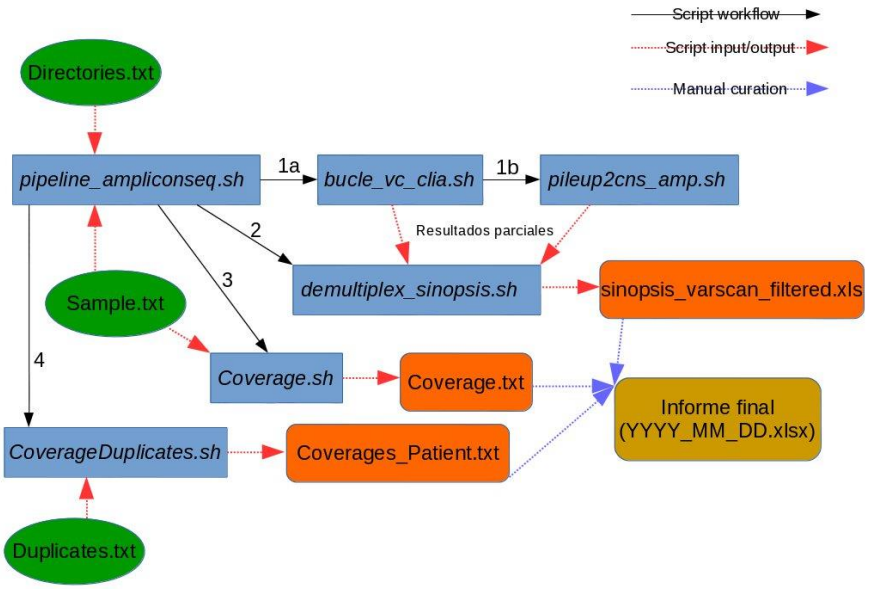

Fig. 5.4 Diagrama de blocs del flux dels fitxers

5.2 Disseny de panells d’oligonucleòtids per a PCR multiplexada

Abans del desenvolupament d’aquest projecte el disseny dels amplicons es duia a terme en cinc passos:

- Especificació de les regions d’interès: en primer lloc s’ha de veure amb quin tipus de gen s’està treballant per tal de considerar-ne les regions d’interès.
  Com ja s’ha esmentat anteriorment, els càncers en òrgans determinats tenen una sèrie de gens associats de forma que, sabent l’òrgan afectat, es poden determinar els gens d’interès que possiblement mostrin mutacions que hagin produït aquest tipus de càncer en concret o que contribueixin al seu desenvolupament. Coneixent quins són aquests gens, es pot buscar a la bases de dades de Cosmic [27] en quines posicions de la seqüència es produeixen les mutacions (en alguns gens es mostren molt localitzades i en altres no).

- Obtenció de la seqüència: mitjançant la base de dades de UCSC (UCSC Genome Browser) [24] s'obté la seqüència del gen. En aquest web es poden trobar les seqüències conegudes de tots els gens de diferents espècies, així com les seves múltiples isoformes observades a la població. Serà necessari conèixer quina d'aquestes variants és la estàndard mitjançant la base de dades que proporciona el Comitè de Nomenclatura de Gens de HUGO (Human Genome Organisation) [26].

  Com es pot observar a la Fig. 5.5, el gen que s'utilitza d'exemple, BRCA1, presenta diverses isoformes però l'estandaritzada és la NM_007294.

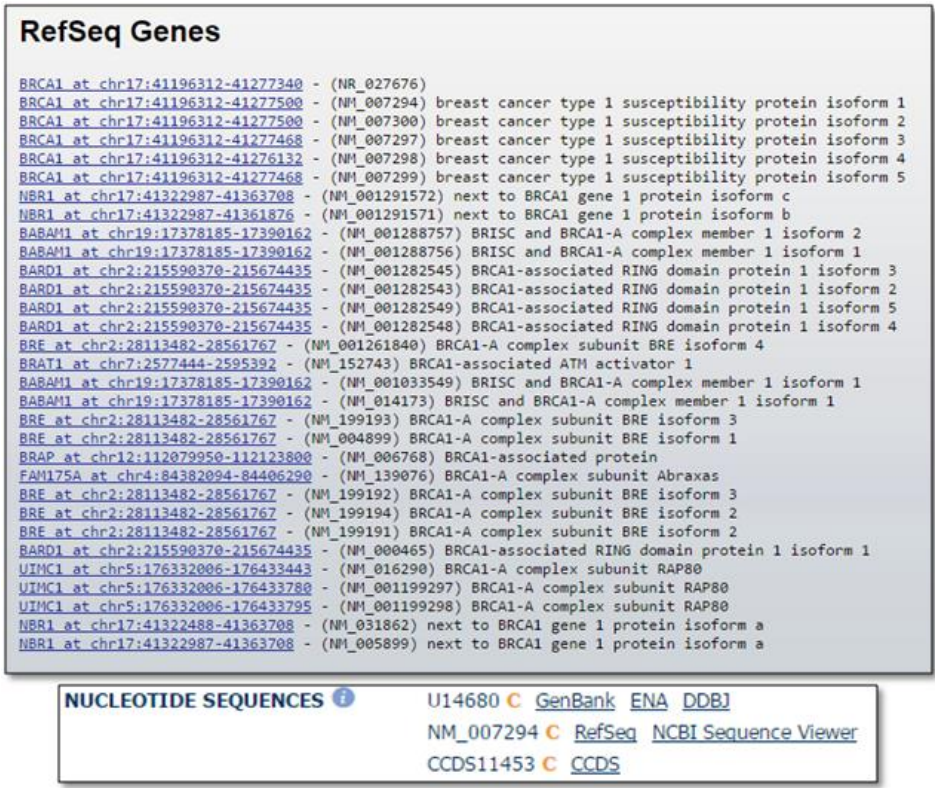

Fig. 5.5 Exemple de les variants que presenta el gen BRCA1 a la base de dades de UCSC (UCSC Genome Browser) i imatge presa de HUGO que determina quina és la estandaritzada.

- Fragmentació de la seqüència: la seqüència obtinguda es presenta de forma de text en fragments que inclouen un exó en majúscula i un cert nombre de bases (a determinar per l'usuari) dels introns que l'envolten. Per tal de dissenyar els oligonucleòtids s'ha de fragmentar aquesta regió en seqüències de mida determinada.

  Per determinar-ne la longitud es realitzen els càlculs pertinents, tenint en compte les limitacions de les màquines i el software que intervenen en la seqüenciació d'ADN.

Tant els càlculs com la fragmentació es realitzen de forma manual, és a dir, es calcula quina ha de ser la longitud dels fragments que es presentaran en majúscula segons el format que requereix el programa Assay Designer com a entrada. De la mateixa forma es determina la longitud dels segments de seqüència que es donaran en minúscula.
Fins ara, aquesta divisió es fa comptant les bases amb l'ajuda del comptador de caràcters de Word, es canvia a majúscula o minúscula manualment i s'introdueixen els caràcters "[" , "/-" i "]" allà on calgui per tal de proporcionar al programa el fitxer d'entrada en el format requerit.

Aquesta tasca pot ser molt feixuga tenint en compte l'elevat nombre de bases nitrogenades que intervenen en tot un gen. A la Fig. 5.6 es pot veure el nombre de bases dels dos primers exons del gen BRCA1, que són molt curts en comparació amb altres exons del mateix gen.

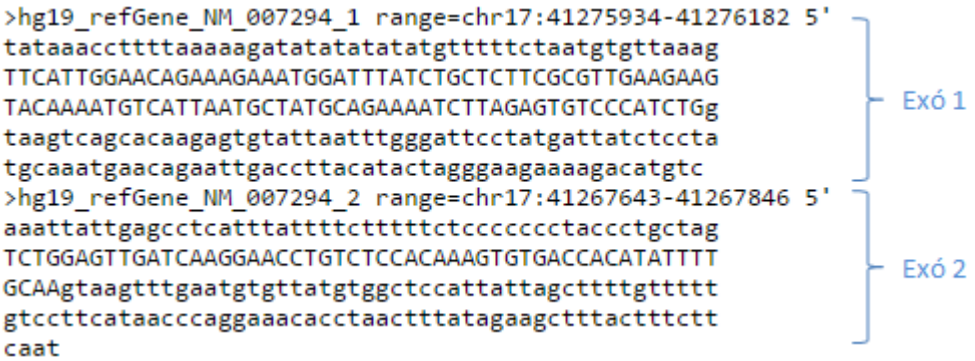

Fig. 5.6 Imatge de la base de dades de UCSC on es mostren les bases que formen els dos primers exons del gen BRCA1

- Disseny d'oligonucleòtids i multiplexatge: els fragments obtinguts s'introdueixen a un programa anomenat Assay Designer, que busca les opcions d'oligonucleòtids més adients per tal de complir els requeriments descrits anteriorment per a una PCR multiplexada. A més, aquest programa agrupa aquells oligonucleòtids que són compatibles en grups anomenats pous o wells.

  És important tenir en compte que la majoria d'amplicons que es dissenyen en regions successives del genoma es superposen per assegurar que tot l'exó quedi ben cobert.

- Obtenció de coordenades: un cop es disposa de les seqüències d'ADN que formaran els amplicons serà necessari conèixer la seva localització. Per a aquest fi s'utlitzarà el web https://genome.ucsc.edu/ que disposa d'una eina, BLAT Seach Genome, que retorna les coordenades dels fragments de seqüència proporcionats

per l'usuari. Aquests valors es copiaran en un nou fitxer en format bed que s'utilitzarà a l'script de la pipeline per alinear-ho contra el genoma humà.

4.3.2 Funcionament dels amplicons

Després de córrer l'scrip Coverage.sh s'obté un fitxer de text amb les dades de cobertura (coverage) dels amplicons. En aquest arxiu es mostra el nombre de lectures (reads) que el MiSeq ha realitzat en una mateixa regió, és a dir, quants cops ha trobat la seqüència amplificada.

Aquests valors reflecteixen l'efectivitat de cada amplicó i serveixen per detectar quines parelles d'oligonucleòtids han funcionat correctament.

El defecte d'aquest sistema és que el programa SAMtools només té en compte la mesura més elevada de cada regió i, per tant, les dades que s'obtenen no són reals. Això és degut a que la majoria d'amplicons estan solapats amb els amplicons veïns i, com que el programa no és capaç de detectar-ho, retorna el valor dels punts de millor cobertura, aquells que estan superposats.

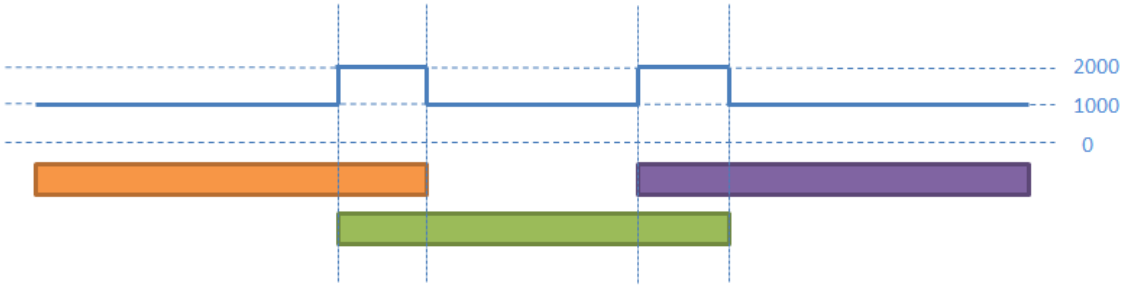

Fig. 4.14 Esquema explicatiu del problema observat. Cada amplicó es representa amb un color diferent i el perfil superior indica el recompte de lectures en un cas hipotètic

A la Fig. 4.14 es pot veure un exemple d'aquest inconvenient: els tres amplicons tenen una cobertura real de 1000 però, quan es solapen, se sumen les lectures i el resultat és 2000. Així doncs, les dades que s'obtindrien serien de 2000 lectures per a cada amplicó, xifres molt allunyades de la realitat.

De la mateixa manera es donen casos en que un amplicó que no funciona sembla tenir una bona cobertura per aquest fet. Això dificulta molt la detecció de parelles d'oligonucleòtids amb mal funcionament, fet que comporta una conseqüència negativa important: perdre mutacions. Com que la regió en qüestió es considera correctament coberta, no detectar mutacions en dit segment del genoma fa pensar que en aquell tram

no hi ha cap anomalia. En realitat pot tractar-se d'un cas amb una mutació important que no serà detectada i tampoc no es plantejarà l'opció de que el motiu sigui un problema dels amplicons, perquè segons el fitxer de cobertura fucionen bé.

Així doncs, es pot afirmar que el procediment actual per al disseny d'amplicons és lent i la forma d'avaluar el funcionament d'aquests no és gaire efectiva. S'haurà de trobar un sistema per poder obtenir dades més acurades.

6 Proposta de solucions

En aquest apartat l'objectiu serà identificar els inconvenients o punts febles que presenta el sistema actual per tal de poder buscar alternatives i, d'entre aquestes, trobar la solució òptima als problemes plantejats.

6.1 Punts febles del sistema actual

Tal i com s'ha vist en apartats anteriors el sistema actual consta de molts passos en els quals intervenen diferents persones del laboratori. Així doncs, aquest treball es centrarà només en la millora d'aquells punts del procés que depenguin de l'equip bioinformàtic, més concretament en el disseny d'amplicons i la verificació del seu correcte funcionament (Fig. 6.1).

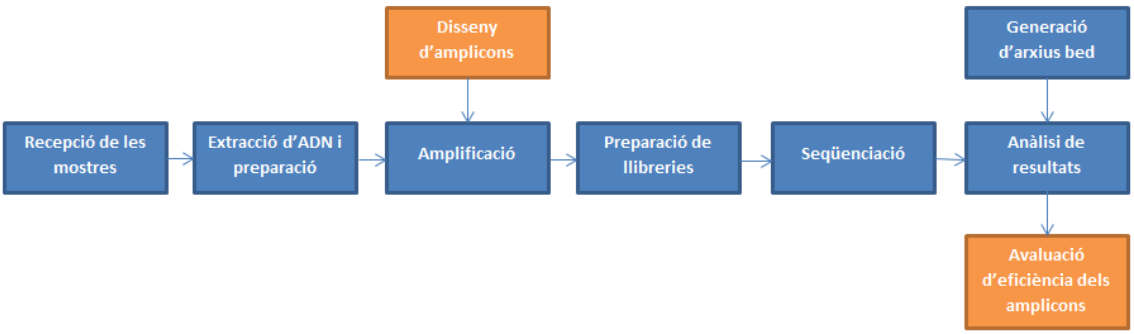

Fig. 6.1 Esquema de blocs del procediment al laboratori, destacant els punts que s'intentaran millorar

6.1.1 Disseny d'amplicons

Prèviament a la realització d'aquest projecte el procediment per a dissenyar oligonucleòtids era llarg i molt repetitiu. Cal recordar que la divisió en subseqüències de cada gen es fa a mà en la seva totalitat, és a dir, es calcula la longitud que han de tenir aquests fragments de la seqüència completa i es procedeix a comptar d'una en una les posicions que formaran part del segment.

Per tenir una idea del temps que això suposa es pot plantejar un exemple: el gen BRCA1 (a la isoforma que es considera estàndard) consta de 23 exons que sumen aproximadament 9000 bases d'interès. Si es considera que generalment les divisions realitzades són de 35 bases i que suposa uns 35 segons preparar una subseqüència amb aquest procediment, s'obté la xifra de 9000 segons, 2,5 hores, per preparar les seqüències d'un sol gen sense comptar amb aturades o errors molt típics de processos d'aquest tipus.

Així doncs, sembla evident que aquest punt milloraria substancialment amb l'ús d'alguna eina per tal d'automatitzar la fragmentació.

6.1.2 Avaluació de l'efectivitat dels amplicons

Per altra banda, hi ha un problema per tal de saber si les dades de cobertura són reals per a cada amplicó. Com s'ha comentat anteriorment, el fet de que els amplicons es solapin provoca que les dades de cobertura que s'obtenen de la pipeline siguin poc acurades, de forma que s'emmascaren dades. És molt difícil saber si una parella d'oligonucleòtids funciona i es sobreestima el coverage total del panell.

6.2 Anàlisi alternatives

Un cop identificats els inconvenients del procés cal proposar solucions per als dos problemes principals que s'han plantejat.

6.2.1 Fragmentació de la seqüència

Sembla evident que la solució més lògica en aquest punt és dissenyar un programa capaç de fragmentar, amb la longitud especificada per l'usuari, les seqüències de cada exó. Així doncs el que caldrà veure són les diferents alternatives de llenguatges de programació de les que es disposa.

En primer lloc serà útil disposar d'un entorn de desenvolupament integrat per tal de facilitar la tasca de disseny del software. Entre l'ampli ventall de IDE's disponibles s'ha considerat oportú triar Microsoft Visual Studio per al disseny d'aquesta eina per diversos motius:

- Ofereix múltiples opcions de llenguatges de programació
- Està dissenyat per a sistemes operatius Windows (els ordinadors del laboratori tenen aquest sistema operatiu)
- Permet crear aplicacions que es comuniquin entre estacions de treball i pàgines web. Aquest aspecte pot ser útil per a possibles millores futures d'aquest programa.
- Té un ús senzill que permet crear interfícies molt fàcilment amb un resultat estètic correcte.

El següent pas consistirà en triar el llenguatge de programació que millor s'adapti als objectius que es volen complir.

Com que una de les premisses d'aquest treball és la de conèixer un nou llenguatge, es descarten en un inici tant Python com C donat que són els que s'han treballat a la universitat. Així doncs les opcions restants són:

- Visual Basic .NET
- F#
- Java
- Ruby
- PHP

6.2.1.1 Visual Basic .NET

Aquest llenguatge de programació orientat a objectes és producte de l'actualització de Visual Basic, un llenguatge dissenyat per Windows.

Els compiladors de Visual Basic generen un codi que requereix de llibreries d'enllaç dinàmic (DLL) per a funcionar. Aquests DLL són arxius de codi executable que es carreguen sota demanda d'un programa per par del sistema operatiu. Aquest fet aporta diversos avantatges com la reducció de la mida dels arxius executables o una gran flexibilitat per a realitzar canvis quan apareguin petits errors.

A més a més, moltes aplicacions de Microsoft Office el porten integrat i és el llenguatge que s'utilitza a l'hora de programar macros d'Excel.

6.2.1.2 F# (F Sharp)

F Sharp és un llenguatge de programació de codi obert que combina la programació funcional amb la disciplina imperativa i l'orientada a objectes. Un dels aspectes positius d'aquest llenguatge és que no cal declarar els tipus de les variables sinó que són deduïts pel compilador.

Proporciona un estructura única de dades integrades directament a la sintaxi del propi llenguatge de forma que permet reduir el temps de desenvolupament d'aplicacions així com la llargada del programa.

6.2.1.3 Java

Aquest és un llenguatge concurrent orientat a objectes que es va dissenyar específicament per tenir el mínim de dependències d'implementació com fos possible. D'aquesta manera es pretenia permetre als desenvolupadors escriure el programa un sol

cop per poder-lo executar després en qualsevol dispositiu, de forma que el codi no haurà de ser recompilat.

Un dels inconvenients és la redundància que presenta en comparació amb altres llenguatges deguda, entre altres factors, a les freqüents declaracions de tipus i les conversions manuals.

Actualment és un dels llenguatges de programació més utilitzats i està present en molts programes i aplicacions.

6.2.1.4 Ruby

Ruby és un llenguatge capaç d'analitzar i executar altres programes que, a diferència dels compiladors, únicament realitza la traducció del codi a mesura que és necessari i generalment no guarda el resultat de dita traducció.

Aquest fet permet produir els mateixos resultats en sistemes molts diferents (com per exemple, un ordinador i una consola de videojocs) però a canvi, presenta menys velocitat que els compiladors a causa de la necessitat d'anar traduint el programa mentre que s'està executant.

6.2.1.5 PHP

Actualment considerat un dels llenguatges més flexibles i potents, PHP es va dissenyar inicialment per al desenvolupament de llocs web de contingut dinàmic.

El codi s'interpreta per un servidor web amb un mòdul de processament específic que regera la pàgina web resultant.

Un dels inconvenient d'aquest llenguatge és el funcionament relativament lent que presenta en comparació amb altres llenguatges, tot i que aquest problema es pot minimitzar amb l'ús de memòries cau.

6.2.2 Cobertura real dels amplicons

Els amplicons presenten una cobertura uniforme al llarg de tota la regió, tal i com es pot veure a la Fig. 6.2. En aquesta imatge es pot observar sota la seqüència d'un tram del genoma, un amplicó anomenat KRAS_A59T_1st_KRAS_A59T_2nd que té les mateixes coordenades que les lectures resultants del seqüenciador (franges grises). Com es pot observar, el gràfic de barres superior indica que el total de lectures ha estat de 5743 i s'ha mantingut constant en tota la longitud de l'insert de l'amplicó.

Així doncs, només caldrà obtenir un punt de cada amplicó que no es trobi solapat amb cap dels amplicons veïns i llegir allà la cobertura per tal d'obtenir el valor real de coverage de cadascun.

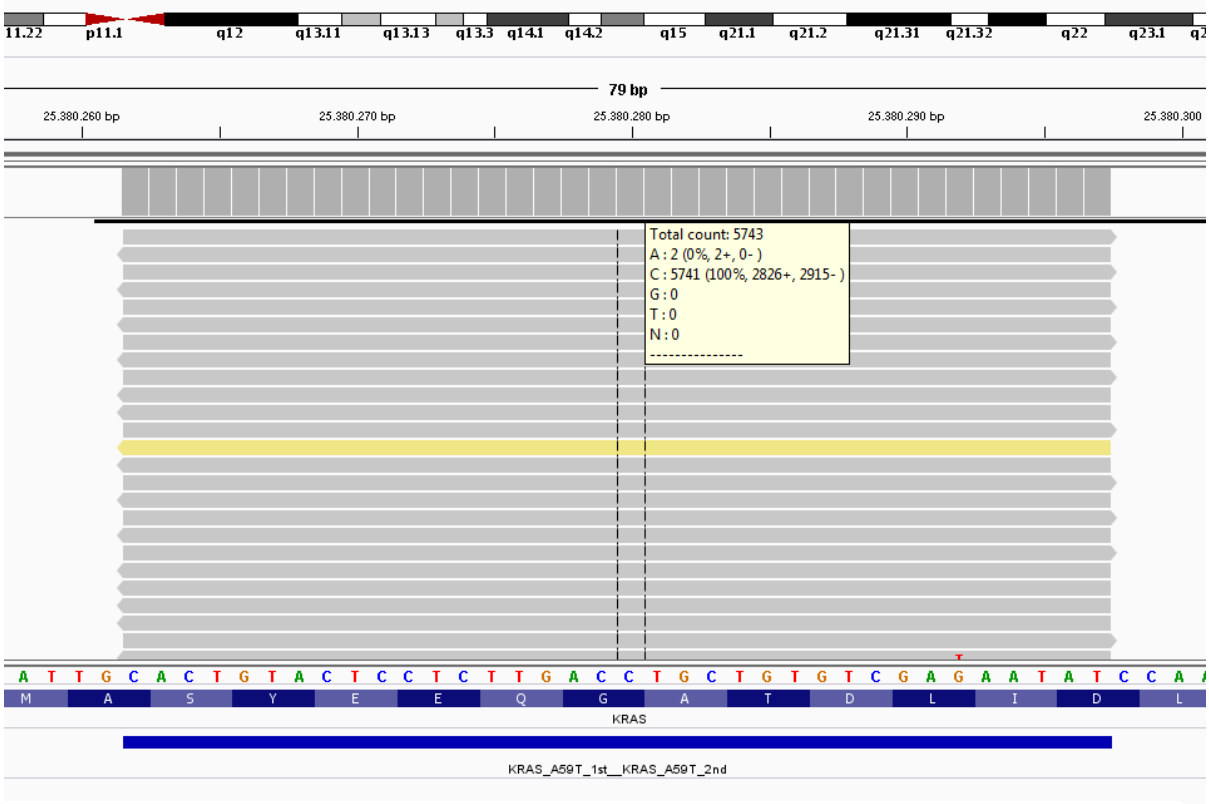

Fig. 6.2 Imatge d'IGV on es mostra la distribució uniforme de lectures quan es tracta d'un amplicó sense solapaments

Per a aquest propòsit hi ha dues possibles solucions:

- Trobar el punt mig de cada amplicó: generalment els amplicons es solapen entre ells al principi i/o al final de manera que, generalment, el punt mig de cadascun d'ells es troba a una zona coberta únicament per aquella parella d'oligonucleòtids.
- Separar el bed original en un bed de regions úniques i un altre de regions repetides: d'aquesta forma només caldrà fixar-se en el bed de regions úniques per veure la cobertura real de cada amplicó.

Totes dues opcions donen solució al problema que es planteja però caldrà triar la més convenient.

El cas de trobar el punt mig sembla l'opció més ràpida i senzilla però amb l'inconvenient que existeix la possibilitat que el solapament dels amplicons no sigui igual en tots els casos, de manera que no es pot assegurar que el punt mig correspongui a una regió

coberta per un sol amplicó, sobretot en els gens on intervenen amplicons dissenyats fa anys.

Per altra banda, l'opció de separar el bed original en dos beds diferents suposarà el desenvolupament d'alguna eina més elaborada però oferirà la seguretat de donar els valors correctes. A més a més es podrà comptar amb la informació referent per a les regions cobertes per més d'un amplicó que també pot ser útil per a veure el funcionament general del panell.

Així doncs sembla que la millor opció serà obtenir dos arxius diferenciats a partir del bed original, que és aquell que conté les coordenades dels oligonucleòtids.

Un cop presa aquesta decisió caldrà trobar la forma més adient per a l'obtenció del bed final. Les dues opcions que es plantegen inicialment són: fer un programa que separi les seqüències o fer ús de les macros d'Excel.

Com que el l'eina per a fragmentar seqüències que es dissenya en aquest treball ja és un programa, s'ha decidit utilitzar el recurs de les macros que proporciona Excel per tal d'adquirir més coneixements al llarg del desenvolupament del treball. A més a més, com que el format bed és un format de text amb files i columnes, es podrà obrir amb Excel sense dificultat i només caldrà guardar els resultats en format de text delimitat per tabulacions.

6.2.2.1 Macros d'Excel

Les macros d'Excel són sèries d'instruccions que s'emmagatzemen per a poder-se executar de forma seqüencial mitjançant una ordre d'execució de forma que permet l'automatització de tasques repetitives.

Segons aquesta definició es pot veure que aquest recurs encaixa perfectament amb les necessitats que es volen satisfer amb aquesta eina i per tant es prendrà la decisió de realitzar una macro d'Excel per a obtenir el fitxer desitjat.

El fet que l'arxiu bed de coordenades es pugui obrir i editar des del propi Excel també influencia positivament en la presa de la decisió.

Només farà falta un pas final que consistirà en aplicar tres ordres des de la línia de comandament de PuTTY a l'arxiu resultant després d'aplicar la macro per tal d'obtenir el bed final. Aquestes ordres comptaran amb el suport del paquet d'eines "*bedtools*" i es guardaran en un arxiu de text per tal que només calgui copiar-les i executar-les.

6.2.2.2 Bedtools

Bedtools és un paquet d'eines que s'utilitzen sovint en el camp de la bioinformàtica per a tractar dades incloses en fitxers de format bed amb operacions a la línia de comandament.

Simplement caldrà descarregar de forma gratuïta el paquet de software de bedtools per tal de poder realitzar diverses operacions com, per exemple, trobar les regions comuns entre fitxers o unir totes les regions de diversos fitxers.

Per tant, només serà necessari disposar del bed que contingui les coordenades dels oligonucleòtids (que s'obté de l'eina BLAT de UCSC [25]), aplicar la macro i copiar les línies de comandament a la consola de PuTTY per tal d'executar-les. D'aquesta manera s'obtindran un bed de regions úniques i un altre de regions repetides o solapades de forma senzilla i ràpida.

6.3 Decisió final

Després de l'anàlisi d'alternatives, sembla que l'opció de crear una macro completa d'Excel juntament amb tres línies de comandament, suposarà una solució senzilla i efectiva per disposar de dos beds que solucionin el problema de càlcul de cobertura plantejat sense perdre informació. A més a més, permetran aprofundir en els coneixements d'Excel (un programa molt útil per a molts aspectes de l'enginyeria) i en l'ús del paquet Bedtools (molt útil en laboratoris de genòmica).

Donat que les macros d'Excel utilitzen el llenguatge Visual Basic, s'ha decidit prendre aquest llenguatge també per al desenvolupament del programa de fragmentació de seqüències.

7 Disseny dels programes

7.1 Programa de fragmentació de seqüències

Tal i com s'ha especificat anteriorment, aquest programa té l'objectiu de fragmentar les seqüències que s'hi introdueixin amb les mides especificades per l'usuari per tal de proporcionar un estalvi de temps en aquest procés

7.1.1 Requisits del programa

Aquest programa haurà de complir amb els requisits que es llisten a continuació:

- Poder-se executar des de qualsevol equip del laboratori
- Tenir un ús senzill i intuïtiu per a persones sense coneixements de programació
- Realitzar càlculs bàsics i fragmentar les seqüències de forma ràpida i amb un sol pas
- Donar flexibilitat a l'hora de decidir la longitud dels fragments que es voldran obtenir
- Proporcionar fragments de seqüència que donin lloc a amplicons superposats per tal d'assegurar la cobertura completa de tota la regió

Aquestes condicions faran que el programa sigui una eina útil de cara a alleugerar la tasca de disseny d'oligonucleòtids.

7.1.2 Descripció del programa

En aquest apartat s'expliquen les dades que requerirà el programa, les seves funcions i la interfície gràfica del mateix. De la mateixa manera es descriurà la forma d'utilitzar aquesta eina.

7.1.2.1 Dades requerides

Per tal d'obtenir els fragments desitjats es requerirà de les següents dades:

- Seqüència completa de l'exó en el que es volen dissenyar els amplicons: es donarà amb el mateix format en el que es troba al web de referència https://genome.ucsc.edu/. Només caldrà copiar-ho i enganxar-ho a l'apartat corresponent
- Mida de l'insert: s'haurà d'especificar la mida aproximada que es desitja per a l'insert, és a dir, la distància entre els dos oligonucleòtids.

- Mida aproximada dels oligonucleòtids: caldrà introduir el nombre de bases desitjades per als oligonucleòtids sense tenir en compte la mida de 10 bases que el programa Assay Design afegeix per defecte a cada oligonucleòtid dissenyat.

Amb aquestes dades d'entrada, el programa fragmentarà la seqüència introduïda en les subseqüències adients per tal que el programa de disseny Assay Designer situï els oligonucleòtids a les regions que compleixin els requisits de l'usuari.

7.1.2.2 Funcions

Les funcions bàsiques del programa són les següents:

- Permet introduir les dades requerides de forma ràpida: simplement cal copiar la seqüència de l'exó i introduir els dos valors de longitud requerits
- Realitza els càlculs bàsics per a determinar la mida dels fragments
- Calcula i mostra en pantalla el nombre d'amplicons que caldran per a cobrir la regió especificada
- Retorna els segments de seqüencia desitjats

7.1.2.3 Interfície gràfica

La introducció de dades per part de l'usuari es realitza a través de la interfície gràfica que es mostra a la Fig. 7.1 .

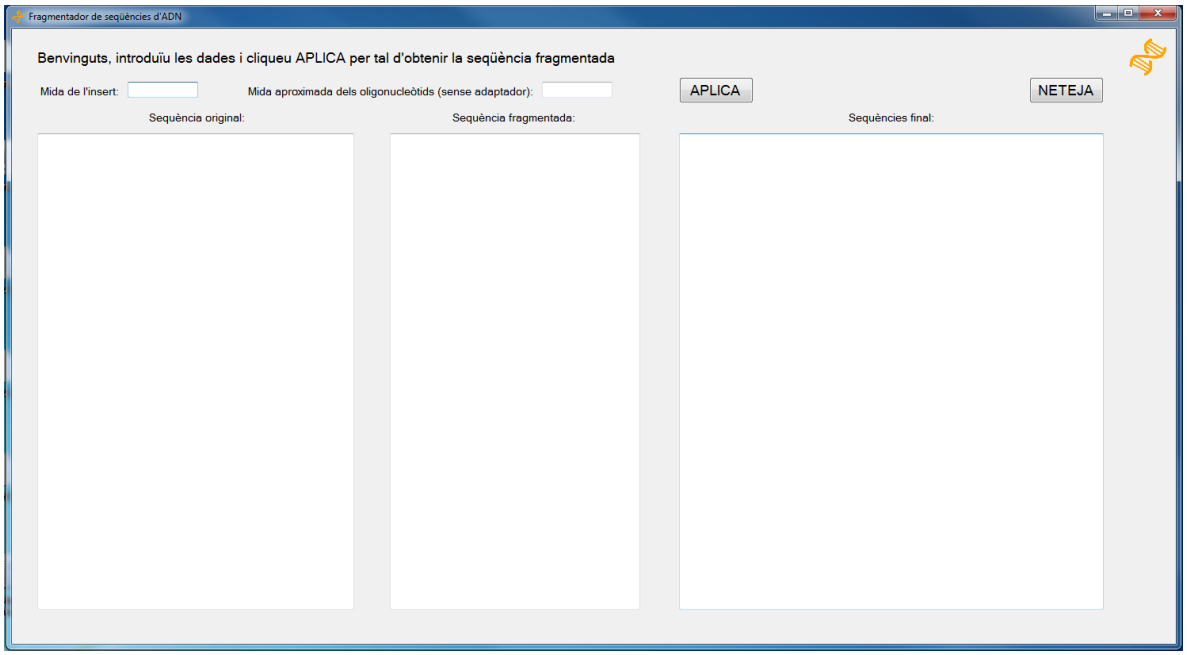

Fig. 7.1 Interfície gràfica

La interfície s'ha dissenyat per permetre la introducció de totes les dades requerides de forma senzilla i intuïtiva. Simplement caldrà prémer el botó que hi apareix per tal de veure el resultat de les seqüències fragmentades a l'apartat corresponent en pantalla (Fig. 7.2).

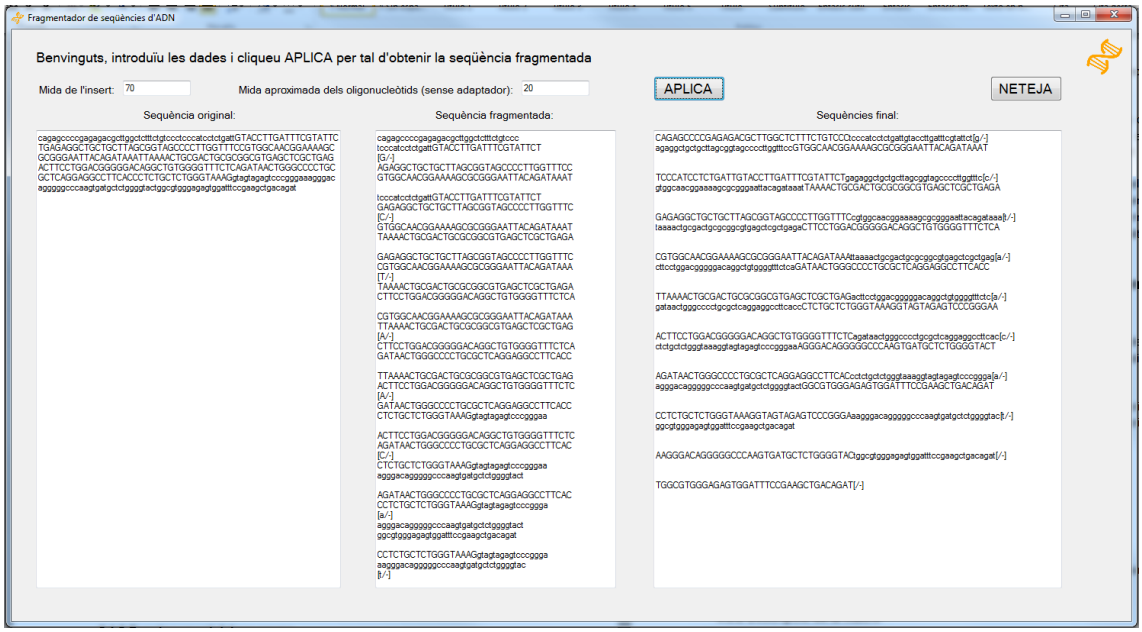

Fig. 7.2 Interfície gràfica amb el resultat donat pel programa

Per passar-ho a l'Assay Design caldrà copiar aquests fragments en un arxiu d'Excel que després es facilitarà al programa per tal d'obtenir el disseny dels oligonucleòtids desitjats.

7.1.3 Funcionament del programa

Per tal d'utilitzar el programa només caldrà tenir-lo instal·lat a l'ordinador i disposar de connexió a internet o disposar de la seqüència de l'exó en qualsevol format de text.

Aquesta seqüència es copia a l'apartat de "Seqüència original" i s'especifiquen les dades requerides a "Mida de l'insert" i "Mida aproximada dels oligonucleòtids". Seguidament s'haurà de clicar "APLICA" i apareixerà en pantalla un missatge amb el nombre d'amplicons que es requeriran i els fragments adients a l'apartat "Seqüència final".

La seqüència final consta de majúscules i minúscules donat que el programa de disseny considera les majúscules com a opcions possibles per a situar-hi els oligonucleòtids i les minúscules com a bases que han de quedar a la zona de l'insert.

També es mostrarà el resultat d'un pas intermig a l'apartat "Seqüència fragmentada" que pot ser útil per a algunes aplicacions. Un exemple és el cas en el que el programa de

disseny no és capaç de trobar oligonucleòtids adients a la regió donada, de forma que l'opció de donar totes les bases del fragment en majúscules proporciona flexibilitat pel programa a l'hora de localitzar els amplicons.

Per tal de realitzar aquestes funcions el codi programa disposa dels següents elements:

- Botons: es disposa de dos botons, clicant-los es realitzen funcions diferents. En el cas del botó "Aplica", s'inicia la fragmentació de les seqüències mentre que el botó "Neteja" s'utilitza per esborrar el contingut de tots els quadres de text.
- TextBox: és una eina que proporciona Visual Studio. Es tracta d'un quadre de text que es pot modificar per l'usuari i pel programa. En aquest cas, es compta amb 5 TextBox que s'utilitzen per permetre la introducció de la seqüència proporcionada per UCSC, la mida dels oligonucleòtids i la mida de l'insert i també per visualitzar les seqüències un cop fragmentades.

En el desenvolupament del programa s'utilitzen dues funcions del llenguatge de programació Visual Basic, de les qual caldrà conèixer el funcionament:

- Funció "Fix": retorna la part entera d'un número. Com a entrada requereix una variable numèrica o una expressió que tingui un nombre com a resultat. En aquest programa s'utilitza per trobar la part entera de la dividir de la mida de l'insert entre 2. Aquest pas és necessari per poder introduir els signes "[", "/-" i "]" al punt mig de les bases de l'insert. Cal recordar que el format de text resultant ha de tenir l'estructura que es mostra a la Fig. 7.3

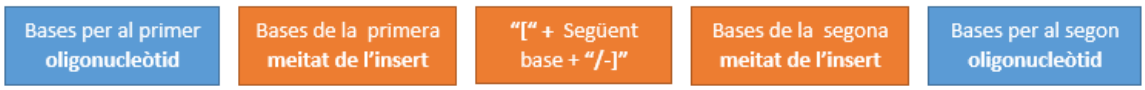

Fig. 7.3 Esquema de l'estructura final dels fragments de seqüència

- Funcio "Mid": retorna una cadena que conté un nombre determinat de caràcters de la cadena d'entrada. Com a entrada es dóna el text del TextBox que conté la seqüència introduïda per l'usuari, un nombre enter que indica a quina posició s'ha de començar a comptar i el nombre de caràcters que caldrà seleccionar. D'aquesta forma es podran obtenir els fragments que calguin en cada pas del programa.
- Funció "Clear": aquesta funció esborra el contingut d'una variable.

Un cop definits els elements de la interfície i les funcions que intervindran, cal conèixer com es desenvolupa el procediment intern del programa. En primer lloc,

l'usuari té dues opcions: netejar el contingut dels quadres de text o iniciar la fragmentació d'una seqüència.

En el primer cas, l'usuari haurà de prémer el botó "Neteja", que aplicarà la funció "Clear" a tots els TextBox.

En cas de prémer el botó "Aplica" s'iniciarà el procés de fragmentació:

- En primer lloc es declaren les variables, entre les quals hi ha dues variables enteres que s'utilitzaran com a comptadors.
- S'eliminen els salts de línia propis de la seqüència que proporciona UCSC.
- Amb l'ajuda d'un comptador i de la funció "Mid" es determinen subseqüències. Del TextBox1 (on l'usuari introdueix la seqüència) es selecciona el nombre de bases requerit pel primer oligonucleòtid i es copia al TextBox2, seguit d'un salt de línia. Després es selecciona la meitat del nombre especificat de bases de l'insert i es realitza la mateixa acció que en el pas anterior. S'afegeix el símbol "[" seguit del caràcter que indiqui el comptador i de "/-]". Després d'un salt de línia, s'afegeix la segona meitat de l'insert i les bases destinades al segon oligonucleòtid, sempre separat per salt de línia. Finalment el comptador es fa retrocedir un nombre de bases que correspon a la suma de la longitud de l'insert i un oligonucleòtid i es repeteix tot el procés. D'aquesta manera el procés generarà seqüències que es solaparan tal i com es veu a la Fig. 7.4.

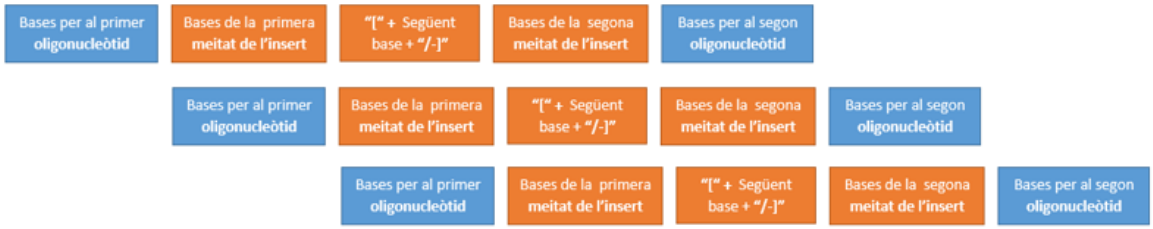

Fig. 7.4 Esquema de superposició de les seqüències

- Seguidament s'utilitzarà el TextBox2 i un altre comptador per generar les seqüències definitives (a la Fig. 7.5 es pot veure el format del TextBox2). El pas final consistirà en posar la primera i l'última línia de cada bloc en majúscula i les línies intermitges en minúscula. Totes les línies del bloc s'ajunten en una sola línia que serà la seqüència definitiva que ja tindrà el format requerit pel programa Assay Design. Aquest pas es pot veure a la Fig. 7.6.

atgcagtgggatctagcatagcgatcgatgcgatc

attgcagtgttcagaCGACATACTGGCATT

ATACGGATTAAAACGACACGTGTCACACAC

[G/-]

TGTCAGCGTTTCAGATCGTGACATGCTAGT

AGCTATACAACACGTAGCCTAGTGCATGTAGCATG

Fig.7.5 Exemple del format de TextBox2

ATGCAGTGGGATCTAGCATAGCGATCGATGCGATC

attgcagtgttcagacgacatactggcatt

atacggattaaaacgacacgtgtcacacac

[g/-]

tgtcagcgtttcagatcgtgacatgctagt

AGCTATACAACACGTAGCCTAGTGCATGTAGCATG

Fig. 7.6 Exemple del format definitiu de les seqüències

7.2 Macro per a l'obtenció de regions úniques i regions repetides

Aquest recurs tindrà com a objectiu l'obtenció de dos beds a partir del bed original dels amplicons. Amb aquesta premissa es definiran els requisits del programa i es procedirà al seu disseny.

Cal tenir en compte que per obtenir el resultat final desitjat s'haurà d'executar una ordre addicional des de la línia de comandament per modificar els arxius que s'obtinguin de la macro. Aquest recurs inclourà una eina anomenada "subtract" continguda en el paquet bedtools.

7.2.1 Requisits

Els requisits que faran d'aquesta macro una eina d'utilitat són els següents:

- Poder-se executar des de qualsevol equip del laboratori
- Tenir un ús senzill i intuïtiu

- Proporcionar dos beds diferents: un de regions úniques i un altre de regions cobertes per més d'un amplicó

7.2.2 Descripció de la macro

Com s'ha fet en el programa per a fragmentar seqüències, en aquest apartat es descriuran les dades que caldran per a executar la macro i les seves funcions, així com la forma d'emprar-la.

7.2.2.1 Dades requerides

Per a utilitzar la macro dissenyada només farà falta disposar del bed amb les coordenades dels oligonucleòtids i del fitxer de la macro en cas que calgui importar-la a Excel.

7.2.2.2 Funcions

La macro d'Excel tindrà dos objectius principals:

- Proporcionar les coordenades de les regions cobertes únicament per un amplicó en un dels fulls de càlcul
- Mostrar en un full de càlcul diferent les coordenades que correspondrien a les regions cobertes com si no hi hagués cap solapament tal i com es pot veure a la Fig. 7.7

Fig. 7.7 Esquema del resultat esperat

7.2.3 Funcionament de la macro

En primer lloc s'haurà d'importar la macro al llibre d'Excel en el que s'estigui treballant. Després d'aquest pas previ caldrà copiar el bed original del que es disposi al primer full de càlcul i executar la macro que realitza els següents passos:

- Assegurar que la coordenada d'inici vagi en primer lloc: l'eina BLAT del portal UCSC proporciona en primer lloc la coordenada inicial de l'oligonucleòtid i després la coordenada final però, en alguns casos, proporciona aquestes dades en l'ordre contrari. En cas que es detecti aquest error, la macro ho corregirà.
- Comprovar que els oligonucleòtids estiguin en l'ordre correcte: en cada parella d'oligonucleòtids n'hi ha un que va en direcció positiva (Forward) i un altre en direcció negativa (Reverse) però, a vegades, l'Assay Design dóna els oligonucleòtids al revés. Si la macro detecta l'error es corregirà.
- Càlcul de les coordenades dels inserts: tal i com es pot veure a la Fig. 7.8, aquestes coordenades seran la coordenada final de l'oligonucleòtid "forward" i la inicial del "reverse".

Fig. 7.8 Esquema de situación de l'insert en relació amb els oligonucleòtids

- Ordre dels inserts: els inserts s'ordenen per la coordenada inicial. Els inserts ordenats es copien a dos fulls de càlcul diferents que s'anomenaran "Regions úniques" i "Regions sense solapament".
- Al full "Regions úniques", en primer lloc es comprova que no hi hagi un cas com el de la Fig. 7.9 i després es calcularan les regions úniques. Per tal de trobar aquestes regions es realitzarà el procés esquematitzat a la Fig. 7.10. Es considera que la coordenada inicial del primer insert no presenta cap solapament i a partir de llavors es comprovarà si la coordenada final té un número superior a la coordenada inicial del següent insert. En cas que això succeeixi, voldrà dir que hi ha solapament i llavors es farà un canvi en les coordenades.

Fig. 7.9 Esquema de cas en que un insert està inclòs en un altre insert

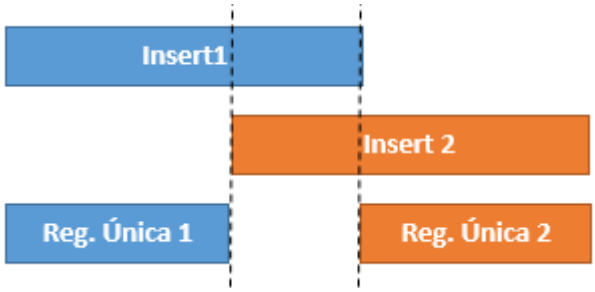

Fig. 7.10 Esquema de coordenades corresponents a les regions úniques 1 i 2

- En una columna addicional es calcula la mida de l'insert per poder descartar els casos de longitud 0 o negativa, que seran aquells en els que no tindrà sentit representar una regió única.
- Al full "Regions sense solapament" s'eliminaran els solapaments canviant les coordenades dels inserts tal i com es mostra a la Fig. 7.11.

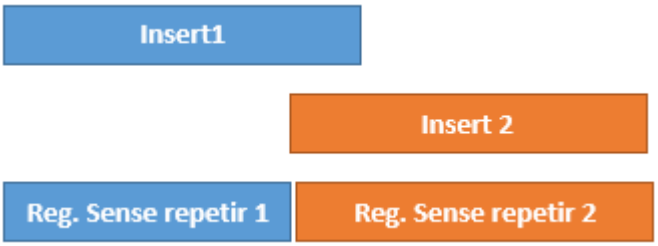

Fig. 7.11 Esquema de coordenades corresponents a les regions no solapades 1 i 2

Finalment només caldrà guardar per separat els fulls de càlcul "Regions úniques" i "Regions sense repeticions" en format de text per posteriorment poder-ho canviar al format bed.

6.2.4 Pas addicional

La funció del pas addicional serà fer una operació semblant a una resta (Fig. 7.12). S'emprarà per a aquest fi l'ús l'eina "subtact" del paquet bedtools.

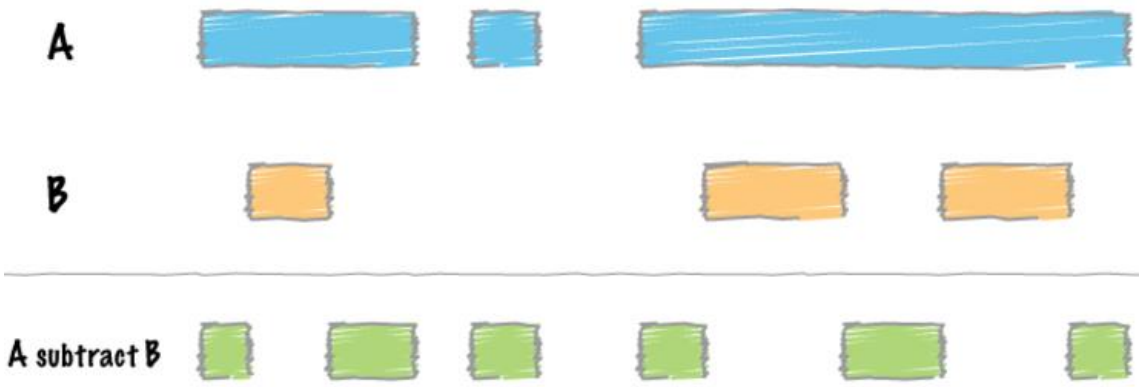

Fig. 7.12 Esquema funcionament de "subtract"

En aquest cas l'eina cerca les coordenades de B que es superposen a A. Quan es produeix aquest fet, la secció superposada s'elimina d'A.

Aquest recurs s'utilitza per a obtenir el bed de les regions cobertes per més d'un amplicó. Per a tal fi es pren el bed de "Regions sense repeticions" que farà el paper d'A a l'exemple anterior. El lloc de B el prendrà el bed de "Regions úniques" de forma que s'obtindrà un bed nou amb les característiques desitjades.

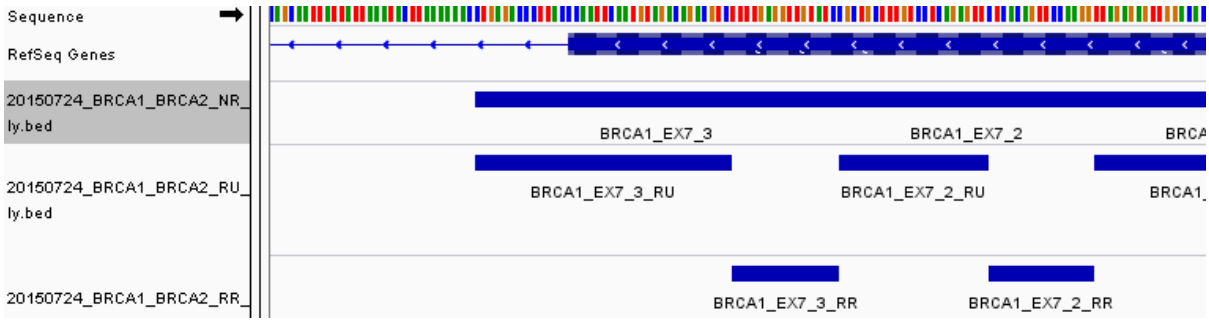

Fig. 7.13 Esquema del resultat esperat

A la Fig. 7.13 es pot veure un exemple d'aquesta aplicació. El bed superior conté les coordenades de la regió com si no hi hagués superposicions. Just a sota trobem el bed amb les coordenades de les regions cobertes únicament per un amplicó i finalment, el bed inferior és el resultat de restar els dos anteriors, és a dir, el que conté les coordenades de totes les regions cobertes per més d'un amplicó.

8 Prova experimental

Per tal de posar en pràctica les aplicacions desenvolupades en aquest treball es procedirà al disseny d'amplicons per a BRCA1, un gen implicat en el càncer de mama i d'ovaris. D'aquesta manera es podrà avaluar l'utilitat de les eines dissenyades al projecte i extreure'n conclusions.

8.1 BRCA1

El gen BRCA1 (breast cancer 1) codifica una proteïna que actua com a supressora tumoral i juga un paper important en la transcripció, recombinació i reparació de l'ADN de ruptures de doble cadena. Segons càlculs recents, aproximadament el 40% de les dones que hereten una mutació nociva d'aquest gen patiran càncer d'ovari abans dels 70 anys. Aquesta xifra augmenta fins al 60% en el cas dels càncers de mama. Les mutacions en aquest gen també poden augmentar el risc de patir càncer de pròstata, trompes de Fal·lopi i pàncrees, entre d'altres [22].

Generalment un gen BRCA1 mutat produeix una proteïna que no funciona correctament i no és capaç d' ajudar a corregir mutacions a altres gens. L'acumulació d'aquests defectes poden permetre a les cèl·lules créixer i dividir-se de forma descontrolada formant un tumor.

Un altre motiu per a triar aquest gen és el fet que recentment es proposen nous tractaments efectius per a càncers que presenten mutacions en BRCA1 i BRCA2, fet que dóna encara més importància a la seva detecció.

Com que BRCA1 presenta mutacions poc localitzades en tota la seqüència, serà un bon exemple per a posar en pràctica les eines desenvolupades al treball.

A la Fig. 8.1 es poden veure dos exemples de gens; el primer (BRCA1) presenta mutacions en tota la seqüència amb un màxim de 23 mutacions detectades en una posició en concret, mentre que el segon (KRAS) té les mutacions clarament localitzades a les posicions 12 i 13 de la seqüència. Aquest exemple és útil per veure la diferència en la distribució de mutacions i entendre per què s'ha triat aquest gen.

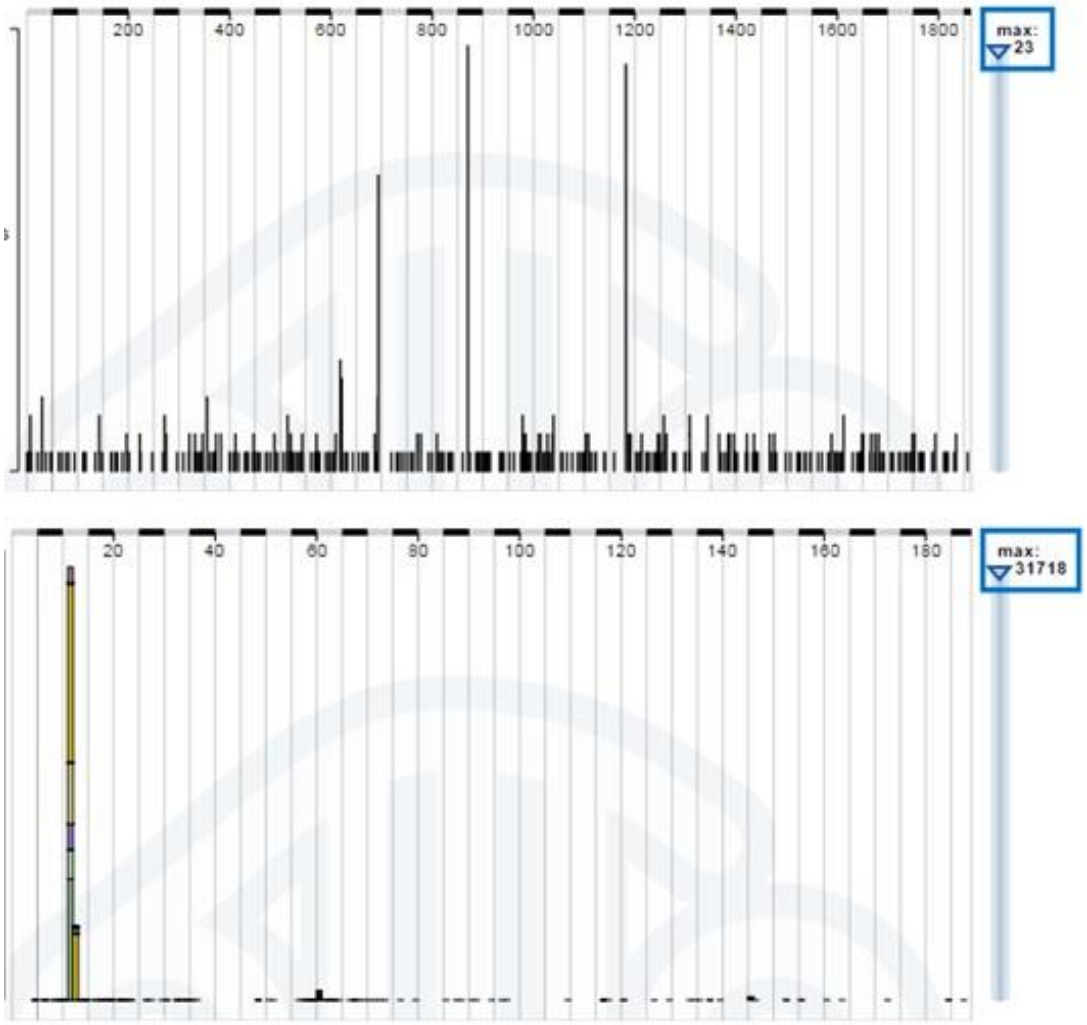

Fig. 8.1 Imatges de la base de dades de Cosmic, que mostren la distribució de mutacions al llarg de la seqüència de BRCA1 i KRAS respectivament

8.2 Obtenció de dades

En primer lloc serà necessari obtenir la seqüència d'ADN d'aquest gen en concret mitjançant el web https://genome.ucsc.edu/. En aquest portal trobem diferents versions del gen, entre les quals haurem de prendre la que es consideri estàndard segons el comitè de nomenclatura de gens HUGO. En aquest cas s'ha vist que la variant d'aquest gen que es considera estàndard és la isoforma 1.

La seqüència obtinguda es troba separada per exons. Cada exó s'introduirà al programa dissenyat per tal d'obtenir els fragments adequats per al format que requereix el software Assay Design. A la Fig. 8.2 es pot veure el format inicial i el resultat que donarà el programa.

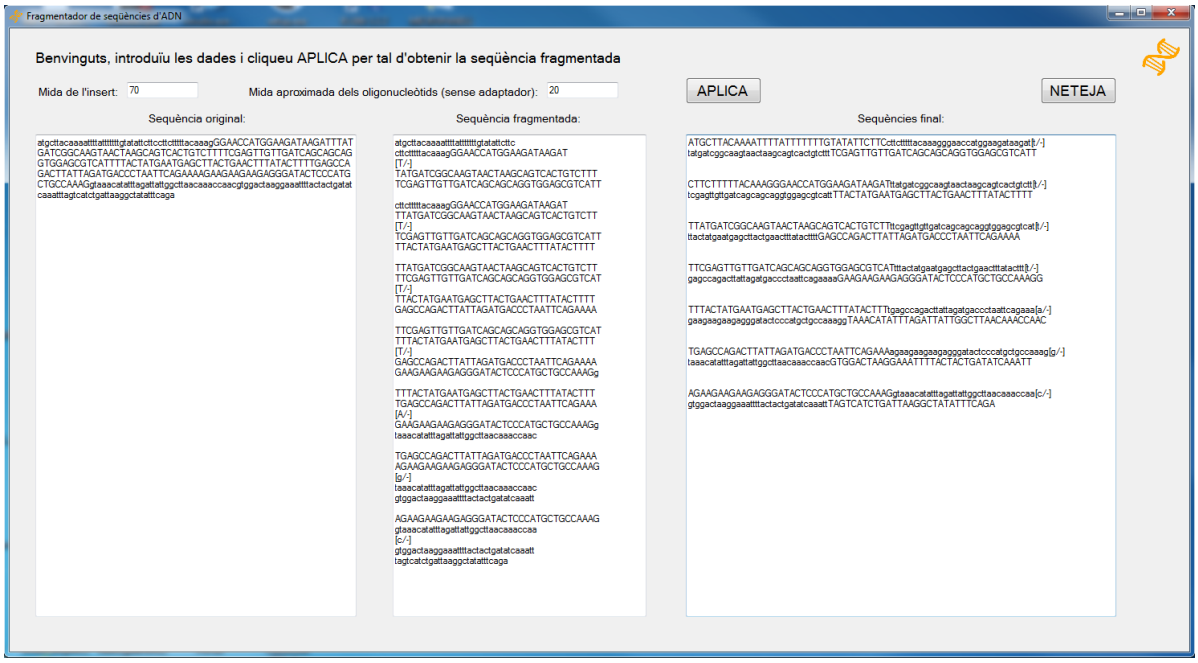

Fig. 8.2 Imatge del resultat que s'obté del programa un cop s'ha fragmentat la seqüència

Com es pot veure a l'annex [Annexos 2.1 i 2.2], dels 23 exons s'han obtingut 175 segments que, després de passar per l'Assay Design han generat 350 oligonucleòtids, recollits a un fitxer. La informació es copia al web https://genome.ucsc.edu/ i, amb l'eina BLAT, s'obtenen les coordenades i es genera el fitxer bed.

```
0         10        20        30        40
17  41276117  41276139  BRCA1_EX2_1_Fwd
17  41276044  41276066  BRCA1_EX2_1_Rev
17  41276008  41276030  BRCA1_EX2_2_Fwd
17  41276081  41276103  BRCA1_EX2_2_Rev
17  41276045  41276067  BRCA1_EX2_3_Fwd
17  41275972  41275994  BRCA1_EX2_3_Rev
17  41267800  41267822  BRCA1_EX3_1_Fwd
17  41267727  41267749  BRCA1_EX3_1_Rev
17  41267691  41267713  BRCA1_EX3_2_Fwd
17  41267764  41267786  BRCA1_EX3_2_Rev
17  41258481  41258503  BRCA1_EX4_1_Fwd
17  41258564  41258586  BRCA1_EX4_1_Rev
17  41258518  41258540  BRCA1_EX4_2_Fwd
17  41258445  41258467  BRCA1_EX4_2_Rev
17  41258409  41258431  BRCA1_EX4_3_Fwd
17  41258482  41258504  BRCA1_EX4_3_Rev
17  41256977  41256999  BRCA1_EX5_1_Fwd
```

Fig. 8.3 Fragment de l'arxiu de format bed on es mostren les coordenades dels oligonucleòtids

Com es pot veure a la Fig. 8.3, a l'arxiu bed es mostren , per ordre, el cromosoma, la posició inicial, la posició final i el nom de l'oligonucleòtid.

A l'arxiu obtingut se li aplicarà la macro per tal d'obtenir el fitxer de regions úniques i el de regions no repetides que posteriorment donaran lloc al fitxer de coordenades definitiu.

Tal i com es pot veure a la Fig. 8.4, el bed anterior queda ben separat en regions que presentaven solapament i regions que estan cobertes únicament per un amplicó. D'aquesta manera només caldrà calcular la cobertura en les regions úniques per conèixer els valors reals de cobertura de cada amplicó.

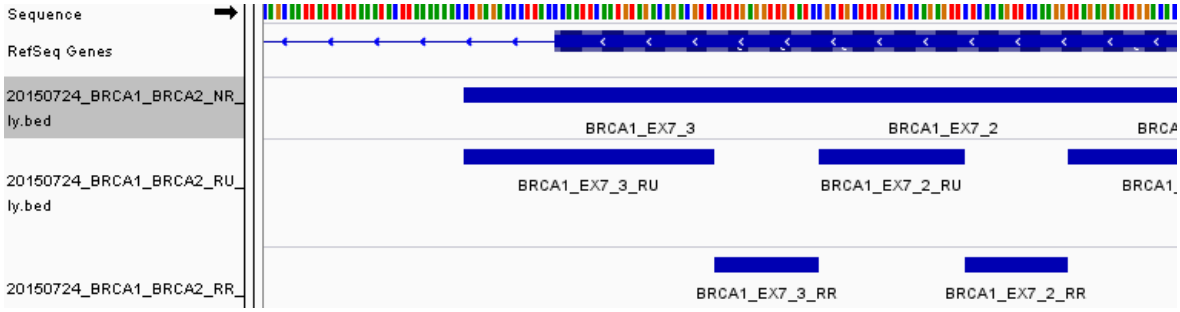

Fig. 8.4 Imatge d'IGV on es mostra en primer lloc el bed de regions sense tenir en compte el solapament, seguidament el bed de regions úniques i, a la part inferior, el bed de regions solapades.

A partir d'aquest punt es procedirà a fer una prova per tal de verificar que els amplicons dissenyats cobreixen les regions esperades i veure quin és el valor de cobertura que s'obté per a cada regió, amb el fi d'identificar amplicons de baix rendiment. Aquest experiment servirà també per comprovar que la separació en regions úniques i repetides que s'ha aplicat al bed és útil per a obtenir dades de cobertura més pròximes a la realitat.

Aquells amplicons que tinguin mala cobertura es podran tenir en compte per a un possible redisseny posterior a la realització del treball.

8.3 Resultats obtinguts

Després de realitzar l'experiment es pot observar amb l'ajuda del software IGV que els amplicons cobreixen la totalitat del gen BRCA1 (Fig. 8.5) però caldrà veure si totes les parelles d'oligonucleòtids ofereixen una cobertura adient de les regions esperades.

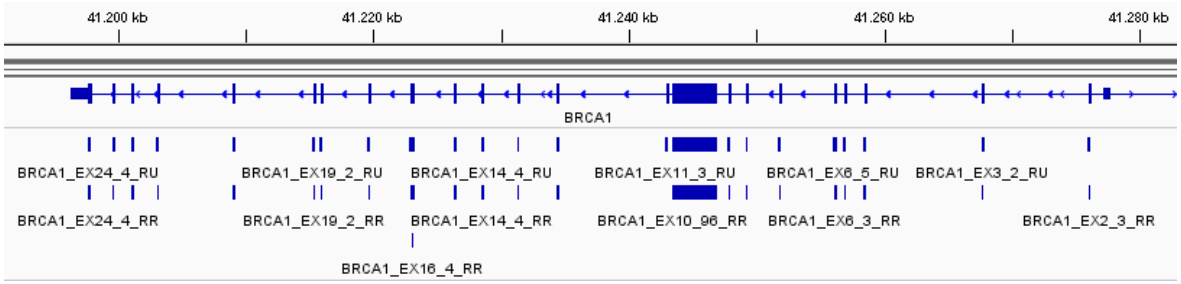

Fig. 8.5 Imatge d'IGV on es pot veure la distribució dels segments d'amplicó resultants després del procés de disseny i aplicación de la macro.

Després de llençar la pipeline i obtenir el fitxer "AmpCoverage.txt" es pot dir que el fitxer de cobertures revela que alguns amplicons han baixat el seu valor de cobertura en comparació amb les dades obtingudes amb el bed original (sense separar regions úniques i repetides). D'aquesta manera, es demostra que hi ha una sèrie d'amplicons que no tenien una cobertura tan alta com es pensava abans de realitzar aquest treball.

Tal i com es pot observar a l'annex [Annex 4], l'experiment de prova s'ha realitzat amb 31 mostres que es representen en columnes separades. El conjunt de mostres s'ha analitzat amb el bed sense modificar i, seguidament, amb el bed resultant després d'aplicar la macro. Els resultats de cobertura es presenten en fulls de càlcul diferenciats.

Per tal de veure el funcionament general de cada amplicó, s'ha realitzat la mitjana aritmètica dels valors de cobertura, que es presenta a la columna anomenada "Mitjana". Observant els valors d'aquesta columna en els dos fulls de càlcul es pot comprovar que són molt més elevats en el cas del bed original. Això és degut al problema que es plantejava a l'inici, el funcionament dels amplicons s'havia sobreestimat degut al solapament d'amplicons. Aquest fet fa pensar que la macro funciona de la forma esperada i que proporciona els valors de cobertura reals corresponent a cada amplicó.

Per tal d'assegurar-ho, es fa la comprovació manual utilitzant l'IGV per visualitzar la situació de les regions úniques i mirant si el valor de cobertura que dóna l'arxiu "AmpCoverages.txt" és correcte. Es pot veure un exemple a la Fig. 8.6. En aquest exemple hi intervenen tres amplicons: BRCA1_EX22_1, BRCA1_EX22_2 i BRCA1_EX22_3. Segons les dades que es veuen a l'annex [Annex 4], els resultats de cobertura amb el bed no modificat pels amplicons esmentats eren 1760, 1760 i 1177 respectivament. En canvi, amb el bed de regions úniques es veu que els valors són 583, 1177 i 0. Amb l'ajuda de la visualització d'IGV es veu que abans d'aquest projecte les dades de cobertura no eren correctes, al produir-se el solapament entre els amplicons BRCA1_EX22_3 i BRCA1_EX22_2, es presentava com a cobertura un valor de 1177 pel primer i 1760 pel segon quan en realitat BRCA1_EX22_3 no funciona i BRCA1_EX22_2 té una cobertura de 1177. Això es deu a que SAMtools va prendre el valor màxim sense

tenir en compte que hi havia un solapament, de forma que, tot i que BRCA1_EX22_3 no funciona, es va prendre el valor del solapament (1177 degut a BRCA1_EX22_2).

Amb aquest exemple queda clar que les coordenades que es donaven pel càlcul no eren correctes per a l'aplicació de SAMtools.

Un cas molt similar es produeix entre BRCA1_EX22_2 i BRCA1_EX22_1, on el valor que es donava per a tots dos en realitat la suma (583+1177=1760).

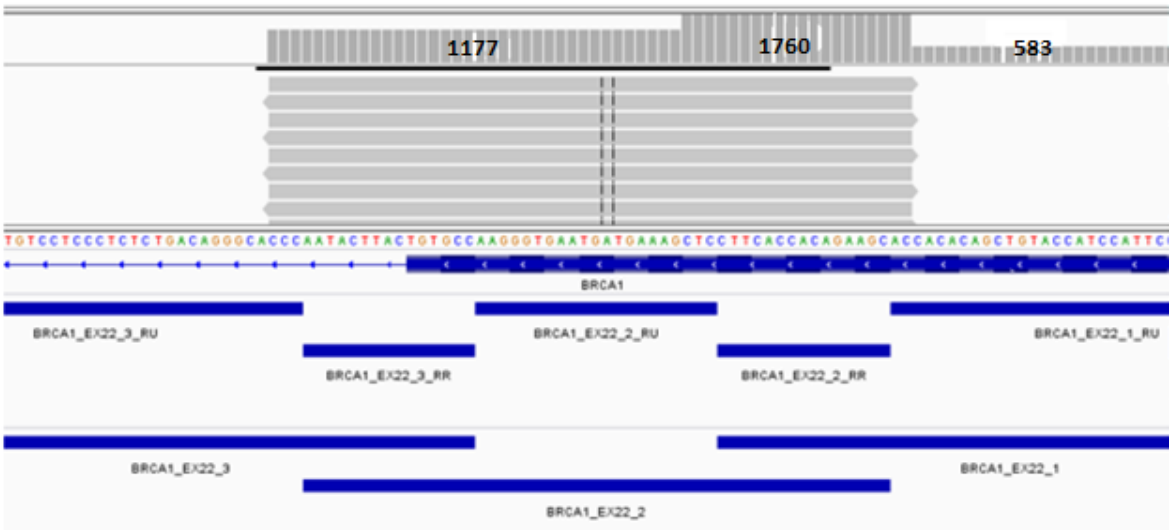

Fig. 8.6 Imatge d'IGV on es pot veure un cas en el que s'havia sobreestimat el funcionament dels amplicons BRCA1_EX22_2 i BRCA1_EX22_1.

El comportament general del conjunt d'amplicons és satisfactori, és a dir, totes les regions queden ben cobertes. Només de 5 regions fallen, probablement per problemes de química que es produeixen sovint en un petit percentatge dels amplicons d'un panell).

Vistos els resultats, es pot afirmar que la divisió realitzada pel programa fragmentador de seqüencies dissenyat és correcta, perquè dóna lloc amplicons que cobreixen la totalitat del gen.

9 Planificació temporal

En aquest apartat s'indicaran les etapes que s'han seguit en la realització d'aquest projecte i el temps dedicat a cadascuna d'elles.

- **Observació i estudi de la metodologia seguida al laboratori de Genòmica del Càncer de l' Institut d'Oncologia de la Vall d'Hebron (25 hores)**: en aquest punt es comença a seguir pas a pas el procés que es porta a terme al laboratori per tal de trobar punts millorables. D'aquesta forma es podran definir els objectius principals del treball.
- **Recerca d'informació referent a la biologia que hi ha darrere del procés (10 hores)**: es fa una recerca de la informació necessària per comprendre les tasques realitzades pel personal del laboratori amb l'objectiu de trobar possibles solucions als problemes observats.
- **Recerca d'antecedents (5 hores)**: s'intenta trobar eines disponibles al mercat o a portals webs que puguin aportar solucions o que presentin un perfil semblant a les eines que es voldrien desenvolupar. En aquesta recerca no es va trobar cap programa o recurs disponible.
- **Anàlisi dels usuaris (10 hores)**: s'analitzen les necessitats de les persones que es beneficiaran del desenvolupament de les eines dissenyades en aquest projecte.
- **Anàlisi dels requeriments de les eines i consideracions a tenir en compte (15 hores)**: en aquesta fase del projecte es defineixen les funcions que hauran de complir les aplicacions per tal de suposar una millora en el procés del laboratori.
- **Introducció a Visual Studio (20 hores)**: es descarrega el programa Visual Studio, es realitzen proves i es segueixen tutorials per tal d'assolir cert domini d'aquesta eina.
- **Introducció al llenguatge de programació Visual Basic .NET (30 hores)**: es segueixen tutorials i es desenvolupen petites aplicacions per poder aprendre les característiques d'un nou llenguatge de programació.
- **Implementació de les aplicacions (120 hores)**: es desenvolupa el programa per a fragmentar seqüències i la macro d'Excel per a l'obtenció de dos beds de regions diferenciades.
- **Detecció d'errors (20 hores)**: un com implementada l'aplicació, es fan proves per identificar defectes i corregir-los.
- **Prova experimental (15 hores)**: en aquest punt s'utilitza una de les eines desenvolupades per dissenyar els oligonucleòtids necessaris per a amplificar el gen BRCA1. Posteriorment, s'empra la macro d'Excel per tal d'obtenir millor càlcul de la cobertura dels productes del disseny.

Aquest pas serà molt important per tal de verificar la utilitat de les aplicacions desenvolupades.

- **Redacció de la memòria (60 hores)**: es redacta la memòria del projecte.

En total, la duració aproximada d'aquest projecte ha estat de 330 hores. Al diagrama de Gantt de la Fig 9.1. es pot veure la planificació al llarg dels sis mesos de durada del projecte.

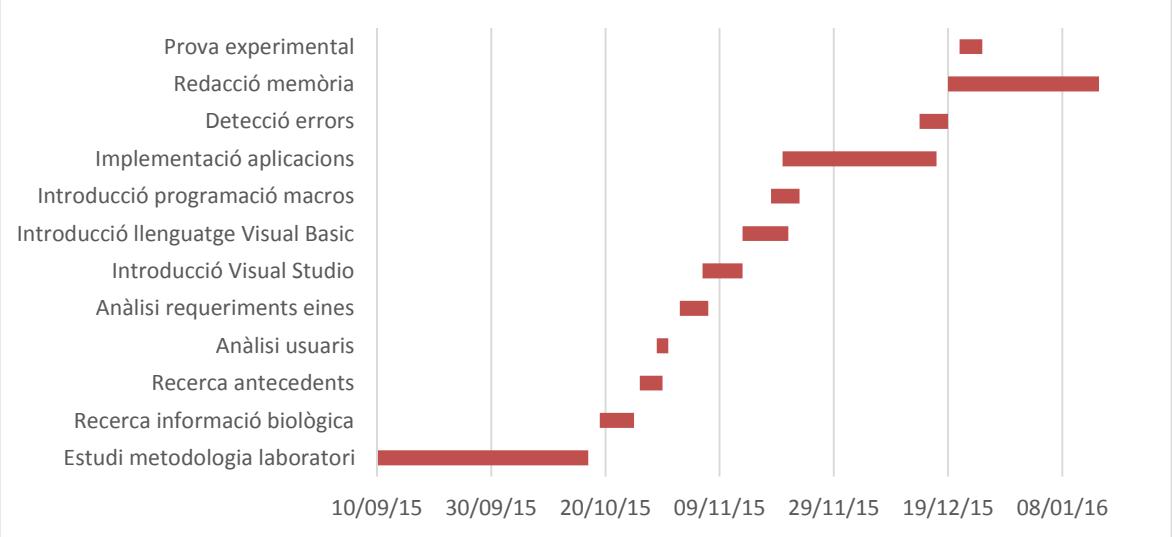

Fig. 9.1 Diagrama de Gantt de la planificació

10 Costos

Els costos d'aquest projecte s'han dividit en tres grups:

- Cost de recursos humans
- Costos directes
- Costos indirectes

10.1 Costos de recursos humans

El projecte ha estat realitzat per una persona que ha realitzat les tasques de disseny i concepció de les aplicacions

Per determinar el cost per hora corresponent a un graduat en Enginyeria en Tecnologies Industrials, s'han pres com a referència les taules salarials de BOE [23]. D'aquesta forma s'ha determinat un cost de 22€/h per a la persona que desenvolupa el procés del treball.

Així doncs, i considerant una durada del projecte de 330 hores, es determina un cost en recursos humans de 7260€.

10.2 Costos directes

En els costos directes (taula 10.1) s'ha inclòs l'adquisició d'un equip informàtic (s'ha calculat el cost de l'equip informàtic del que s'ha disposat per desenvolupar el treball).

En aquest apartat també s'hi haurien d'incloure les llicències dels programes que s'han emprat. Tots els programes emprats en el desenvolupament d'aquest projecte eren lliures, exceptuant el software de les màquines (que ja estava disponible al laboratori abans de la realització d'aquest treball) i el programa Visual Studio. Aquest programa disposa d'una versió de prova però només és vàlida durant un mes, de forma que caldrà adquirir la versió professional, amb un cost de 45$ al mes (41.34€ segons el canvi de monedat actual).

Concepte	*Cost*
Equip informàtic	1821€
Llicència Visual Studio (durant 6 mesos)	248.04€
TOTAL	**2069.04€**

Taula 10.1 Càlcul desglossat de costos directes

10.3 Costos indirectes

En el càlcul de costos indirectes només s'ha tingut en compte el cost de la connexió a internet donat que el càlcul del consum elèctric de l'equip informàtic resulta molt difícil d'estimar. En aquest àmbit, s'ha considerat un cost de 40€ al mes per a la tarifa d'internet, de forma que l'ús d'internet durant aproximadament 6 mesos suposa un total de 240€.

10.4 Cost total

Finalment, i després de calcular els costos de recursos humans, els costos directes i els costos indirectes, es determina que el cost total del projecte és de 9569.04€.

11 Impacte sobre l'entorn

En aquest punt s'estudia breument l'impacte que l'ús de les aplicacions desenvolupades en aquest treball ocasionen en el conjunt de persones, ecosistemes i béns que envolten el projecte.

Aquestes aplicacions tenen com a objectiu agilitzar el disseny d'oligonucleòtids i millorar l'eficiència en el càlcul de cobertura de les regions incloses en els panells d'amplicons.

L'usuari necessita d'un equip informàtic per poder utilitzar les aplicacions, fet que comporta una despesa d'energia elèctrica. Donat que el temps de disseny es minimitza amb l'aplicació de les eines desenvolupades, es pot afirmar que aquesta despesa decreixerà.

Les aplicacions estan enfocades a facilitar tasques al laboratori, fet que es pot considerar un impacte positiu en l'àmbit de la investigació. Per altra banda, la millora en el càlcul de cobertures permet identificar de forma més eficient l'eficiència dels amplicons per tal de poder-los redissenyar per cobrir correctament les regions d'interès del genoma. Aquest fet suposa un aspecte positiu de cara a la millora de tractaments que poden rebre els pacients i suposa un impacte positiu sobre la societat.

12 Conceptes biològics d'interès

Donat que aquest és un treball multidisciplinari, s'ha considerat oportú afegir aquest apartat de nocions bàsiques de biologia essencials per a la comprensió d'alguns processos que apareixen en aquesta memòria.

Les explicacions que es presenten a continuació no pretenen aprofundir en excés en la biologia, sinó donar una idea senzilla i clara d'alguns conceptes per tal de facilitar la lectura d'aquesta memòria.

Aquests conceptes es llisten a continuació:

Àcid nucleic: polímers formats per la repetició de monòmers anomenats nucleòtids que s'uneixen mitjançant enllaços fosfodièster formant llargues cadenes. Emmagatzemen la informació genètica dels organismes vius i són els responsables de la transmissió hereditària. N'hi ha de dos tipus: ADN i ARN.

ADN: àcid nucleic que emmagatzema la informació necessària per a construir components de les cèl·lules com les proteïnes i les molècules d'ARN. Té estructura en doble hèlix i bicatenària (Fig 12.1), és a dir, formada per dues cadenes antiparal·leles amb bases oposades (A amb T i C amb G) tal i com es pot veure a l'esquema de la Fig. 12.2

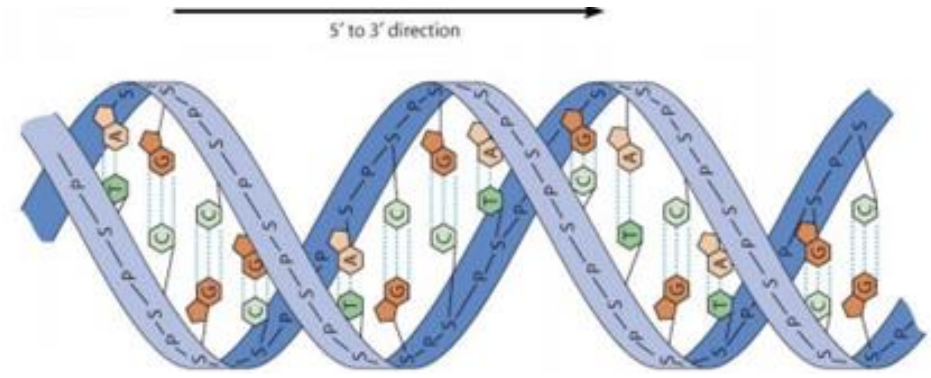

Fig. 12.1 Esquema de l'estructura de l'ADN

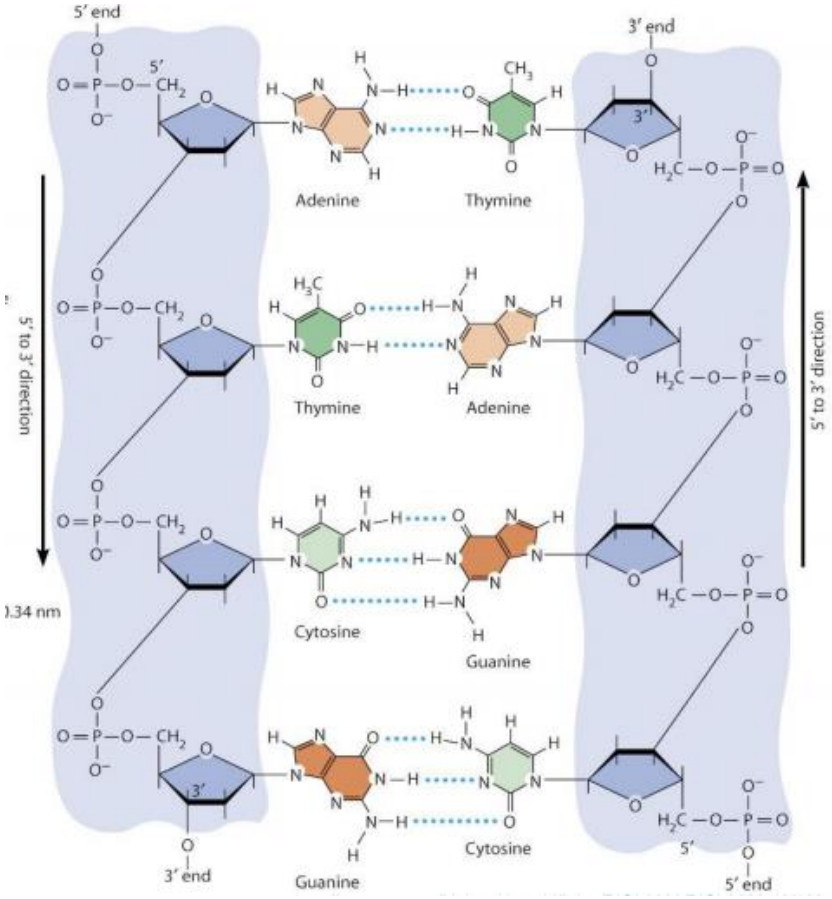

Fig. 12.2 Imatge d'IGV on es pot veure un cas en el que s'havia sobreestimat el funcionament dels amplicons BRCA1_EX22_2 i BRCA1_EX22_1.

Aminoàcid: molècula orgànica que es combina amb altres per a formar proteïnes. A la Fig. 4.2 es pot veure una taula d'equivalència entre triplets de bases i els aminoàcids corresponents.

Amplicó: fragment d'ADN producte de l'amplificació a partir d'una parella d'oligonucleòtids, tal i com s'observa a la Fig. 12.3

Fig. 12.3 Esquema de l'estructura d'un amplicó

ARN: àcid nucleic que intervé en la síntesi de proteïnes. Transmet la informació genètica entre l'ADN que es troba al nucli de la cèl·lula fins al citosol, que és on es sintetitzen les proteïnes. La seva estructura és de cadena simple.

Hi ha diferents tipus d'ARN entre els qual destaquen:

- ARN ribosòmic (ARNr): tradueix l'ARN missatger a aminoàcids.
- ARN missatger (ARNm): es produeix a partir de la transcripció de l'ADN. Transporta la informació genètica des del nucli de la cèl·lula fins al citoplasma, on hi ha l'ARN ribosòmic.
- ARN de transferència (ARNt): té la funció de reconèixer cada triplet de nucleòtids de l'ARN missatger amb l'ARN ribosòmic de forma complementària, és a dir, que ARNm i ARNr tindran la mateixa seqüència però invertida i amb bases complementàries. Per cada triplet d'ARNm que llegeix, s'enllaça un aminoàcid.

Base nitrogenada: compostos orgànics amb dos o més àtoms de nitrogen que són part fonamental de nucleòtids i àcids nucleics. L'ADN està format per quatre bases nitrogenades: adenina (A), la guanina (G), la timina (T), la citosina (C) L' ARN el formen les mateixes bases però amb l' uracil (U) substituint a la timina.

Cebador: cadena d'àcid nucleic que serveix com a punt de partida per a la còpia de l'ADN. Conté un grup hidroxil lliure que forma parelles de bases complementàries a una cadena que serveix de motlle i que actua com a punt d'inici per a l'addició de nous nucleòtids amb l'objectiu de generar una còpia de la cadena motlle.

Enzim: molècules que catalitzen reaccions químiques accelerant el procés.

Exó: regió d'un gen que no es separa durant el procés d'splicing i que, per tant, es manté en el producte d'ARN. És a dir, és la part del gen que conté la informació per a la síntesi de la proteïna. Cada exó codifica una part específica de la proteïna.

Fenotip: característica o tret característic d'un organisme, com la morfologia, propietats bioquímiques o comportament, que es produeix com a conseqüència de l'expressió del genotip en funció d'unes determinades condicions.

Gen: seqüència lineal de nucleòtids d'ADN. És considerat com la unitat d'emmagatzematge d'informació i responsable de la transmissió per herència . Estan constituïts per regions codificants (exons) i regions no codificants (introns) que s'eliminen en el procés d'splicing. Després del procés de transcripció i traducció donarà lloc a una proteïna (Fig. 12.4).

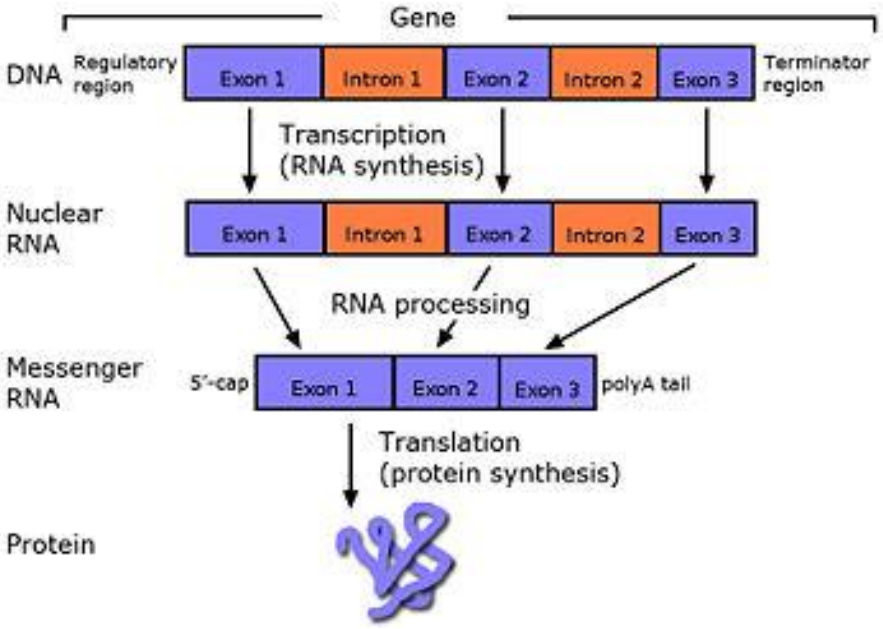

Fig. 12.4 Esquema del procés per arribar a generar proteïnes

Genoma: conjunt de gens continguts als cromosomes.

Genòmica: branca de la biologia dedicada a l'estudi del funcionament, contingut, evolució i origen dels genomes.

Genotip: informació genètica que aporta l'ADN a un organisme en particular.

Insert: regió d'ADN compresa entre dos oligonucleòtids, és a dir, serà la regió amplificada en el procés de PCR.

Intró: regió de l'ADN compresa a la part codificant d'un gen però que no s'arriba a expressa, és a dir, la seva seqüència no s'utilitza quan se sintetitza la corresponent proteïna, tal i com es pot veure a la Fig 12.4.

Isoforma: cadascuna de les diferents formes que pot prendre una proteïna a causa d'un splicing alternatiu del mateix gen.

Mutació: canvi permanent en l'ADN que es pot produir per acció d'agents externs a la cèl·lula o per agents interns com errades en la replicació o reparació de l'ADN.

Nucleòtid: molècules orgàniques que contenen una base nitrogenada, un àcid fosfòric i una pentosa. Són els monòmers dels àcids nucleics.

Oligonucleòtid: fragment curt d'ADN que actua com a cebador en reaccions d'amplificació com la PCR.

Polimerasa: enzim capaç de transcriure o replicar àcids nucleics mitjançant l'addició de nucleòtids lliures a un fragment de cadena simple utilitzant la cadena complementària com a motlle.

Polimorfisme d'un sol nucleòtid: variacions en la seqüència d'ADN que afecten a una base i que es detecten recurrentment en individus de la població.

Splicing: després de la transcripció d'ADN s'obté ARNm immadur que tindrà la mateixa estructura d'exons i introns. El procés d'splicing (Fig. 12.5) és aquell en el qual es suprimeixen els introns i es mantenen únicament els exons, que s'uneixen entre ells.

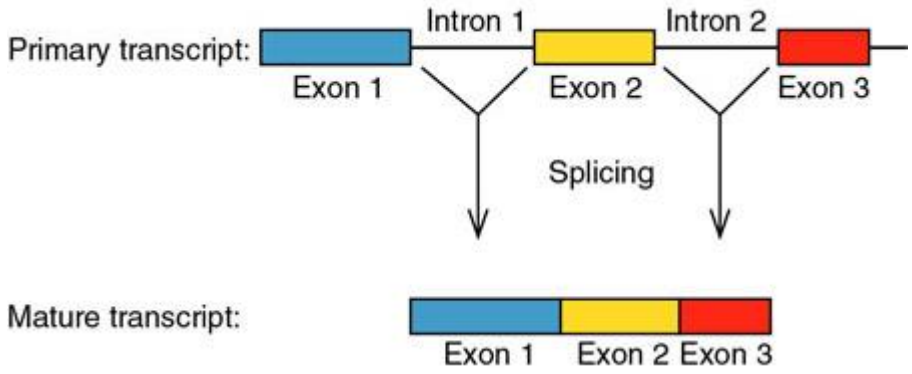

Fig. 12.5 Esquema del procés d'splicing

Taq polimerasa: és un tipus de polimerasa que té l'avantatge de ser termoestable.

Well: anglicisme que s'utilitza per a referir-se als petits recipients on s'introdueix la barreja de reactius.

Conclusions

L'objectiu principal d'aquest treball és desenvolupar eines per tal d'agilitzar el procés de disseny d'amplicons i calcular-ne l'eficiència de forma més acurada. Aquest objectiu s'ha assolit mitjançant un programa i una macro d'Excel .

Després de realitzar una prova experimental, es pot afirmar que el programa proporciona una fragmentació de seqüències adequada, donat que els amplicons resultants han cobert la totalitat del gen escollit proporcionant bons valors de cobertura. Per altra banda, el temps requerit per a obtenir els oligonucleòtids necessaris s'ha reduït substancialment.

Al punt de disseny d'amplicons d'aquesta memòria s'ha plantejat un exemple del temps necessari per a fragmentar tot el gen BRCA1 abans de la realització del projecte, amb un resultat aproximat de 2,5 hores. Després de la prova experimental, es pot dir que el temps necessari és de 23 minuts, el que suposa un estalvi de temps del 85%.

Pel que fa al càlcul de l'eficiència dels amplicons, s'ha pogut veure que la macro dóna solució al problema plantejat. Els càlculs de cobertura que s'obtenen amb el bed resultant de l'aplicació d'aquesta macro són correctes i no presenten errors quan es produeixen solapaments entre amplicons.

Durant aquest treball s'han assolit nous coneixements, sobretot s'ha aprofundit en el coneixement d'un nou llenguatge de programació (VisualBasic.NET), en el desenvolupament de programes amb interfície i en l'ús avançat d'Excel. A més a més s'han assolit coneixements biològics de l'àmbit de la genòmica del càncer.

Per últim, la realització d'aquest projecte ha mostrat la capacitat assolida d'analitzar un procés desconegut en un inici, per tal de trobar-ne punts febles i cercar la millor opció per millorar-los.

Agraïments

M'agradaria expressar el meu agraïment a les persones que han fet possible aquest treball. En especial a Núria Pla i Francesco Mancuso, que han guiat el desenvolupament del projecte i a Ana Vivancos per donar-me l'oportunitat de dur-lo a terme.

També voldria agrair a tots els tècnics del laboratori la paciència que han mostrat per ajudar-me a comprendre la seva feina.

Per últim voldria donar les gràcies a la meva família i amics, en especial a Lorena Torres, per recolzar-me i donar-me un cop de mà en tot el que han pogut.

Bibliografia

Referències bibliogràfiques

[1] **Metzker, M. L.** Sequencing technologies - the next generation. *Nature Reviews. Genetics*, [en línia] http://www.nature.com/nrg/journal/v11/n1/full/nrg2626.html [Consulta:10 setembre 2015]

[2] **Shendure, J.; Ji, H.** Next-generation DNA sequencing. Nat Biotechnol *26* [en línia] http://www.nature.com/nbt/journal/v26/n10/full/nbt1486.html [Consulta: 10 setembre 2015]

[3] **Gualberto, J.** PCR Protocols. Plant Science [en línia] http://www.sciencedirect.com/science/article/pii/S0168945204000846 [Consulta 12 setembre 2015]

[4] **Henegariu, O.; Heerema, N. A.; Dlouhy, S. R.; Vance, G. H.; Vogt, P. H.** Multiplex PCR: Critical parameters and step-by-step protocol.BioTechniques [en línia] http://link.springer.com/article/10.1007%2FBF00928712 [Consulta: 15 setembre 2015]

[5] **Cerami, E.; Gao, J.; Dogrusoz, U.; Gross, B. E.; Sumer, S. O.; Aksoy, B. A., Jacobsen, A.** The cBio Cancer Genomics Portal: An open platform for exploring multidimensional cancer genomics data. Cancer Discover [en línia] http://cancerdiscovery.aacrjournals.org/content/2/5/401 [Consulta: 18 setembre 2015]

[6] **Quinlan, A. R.; Hall, I. M** The BEDTools manual. *Genome* [en línia] https://code.google.com/p/bedtools/downloads/detail?name=BEDTools-User-Manual.v4.pdf&can=2&q= [Consulta: 30 setembre 2015]

[7] **SamTools Team** Sam, T., & Specification, F. The SAM Format Specification [en línia] http://samtools.sourceforge.net/ [Consulta: 30 setembre 2015]

[8] **Li, H.; Handsaker, B.; Wysoker, A.; Fennell, T.; Ruan, J.; Homer, N.; Marth, G.** The Sequence Alignment/Map format and SAMtools. *Bioinformatics* [en línia] http://bioinformatics.oxfordjournals.org/content/25/16/2078.full [Consulta: 30 setembre 2015]

[9] **Xu, G.; Deng, N.; Zhao, Z.; Judeh, T.; Flemington, E.; Zhu, D.** SAMMate: a GUI tool for processing short read alignments in SAM/BAM format [en línia] http://scfbm.biomedcentral.com/articles/10.1186/1751-0473-6-2 [Consulta: 30 setembre]

[10] **Cock, P. J. A; Fields, C. J.; Goto, N.; Heuer, M. L.; Rice, P. M.** The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Research* [en línia] Available from Michael Heuer's profile on Mendeley. [Consulta: 3 octubre 2015]

[11] **Nielsen; R.; Paul; J. S.; Albrechtsen; A.; & Song; Y. S.** Genotype and SNP calling from next-generation sequencing data. *Nature reviews. Genetics*,*12*(6), 443-451. Nature Publishing Group [en línia] http://www.nature.com/nrg/journal/v12/n6/full/nrg2986.html [Consulta: 7 octubre 2015]

[12] **Mackey, A**. *Introducing .NET 4.0 with Visual Studio 2010*. *Apress* . Apress, 2010 [Consulta: 7 novembre 2015]

[13] **Stott, W.; Newkirk, J**. *Visual Studio Team System*. *Source* Addison-Wesley 2005 [Consulta: 8 novembre 2015]

[14] **Moore, A.** *Visual Studio 2010 All-in-One For Dummies*. *Building* Wiley Publising,Inc 2010 [Consulta:11 novembre 2015]

[15] **Jelen, B; Syrstad, T.** *VBA and Macros: Microsoft Excel 2010* Que 2010 [Consulta: 19 novembre 2015]

[16] **Roman, S.; Roman, B. S.** *Writing Excel macros with VBA*. *Cognition* O'Reily Media, Inc 2002 [Consulta: 21 novembre 2015]

[17] **Hanahan, D**. The Hallmarks of Cancer. *Cell* [en línia] http://linkinghub.elsevier.com/retrieve/pii/S0092-8674(00)81683-9\nhttp://www.sciencedirect.com/science/article/B6WSN-4195FC1-5/2/aef1d48431eadea4567b697b1fee0514 [Consulta: 30 novembre 2015]

[18] **Hanahan, D.; Weinberg, R. A**. Hallmarks of cancer: the next generation. *Cell* [en línia] http://www.cell.com/cell/fulltext/S0092-8674(11)00127-9 [Consulta: 2 desembre 2015]

[19] **Deng, C.-X**. BRCA1: cell cycle checkpoint, genetic instability, DNA damage response and cancer evolution. *Nucleic acids research* [en línia] http://nar.oxfordjournals.org/content/34/5/1416 [Consulta: 5 desembre 2015]

[20] **Clark, S. L.; Rodriguez, A. M.; Snyder, R. R.; Hankins, G. D. V.; Boehning, D.** Structure-Function Of The Tumor Suppressor BRCA1. *Computational and Structural Biotechnology Journal*, *1*(1), 1-8. Research Network of Computational and Structural Biotechnology [en línia] Available from Darren Boehning's profile on Mendeley [Consulta: 21 desembre 2015]

[21] **Economist Inteligence Unit (European Parliament)** Living Downstream: A personal exploration of cancer and our environment & European 'environmental prevention' policies [en línia] http://www.env-health.org/IMG/pdf/Living_Downstream_Growth_in_cancer_incidence_in_EU_countries.pdf [Consulta: 23 desembre 2015]

[22] **Instituto Nacional del Cáncer** BRCA1 y BRCA2: *Riesgo de cáncer y pruebas genéticas* [en línia] http://www.cancer.gov/espanol/cancer/causas-prevencion/genetica/hoja-informativa-brca#q2 [Consulta: 24 desembre 2015]

[23] Boletín Oficial del Estado Resolución de 9 de octubre de 2013, de la Dirección General de Empleo, por la que se registra y publica el XVII Convenio colectivo nacional de empresas de ingeniería y oficinas de estudios técnicos. [en línia] https://www.boe.es/boe/dias/2013/10/25/pdfs/BOE-A-2013-11199.pdf [Consulta: 27 desembre 2015]

[24] UCSC Genome Browser: Kent, WJ; Sugnet, CW; Furey, TS; Roskin, KM; Pringle, TH; Zahler, AM; Haussler, D. The human genome browser at UCSC. *Genome Res.* 2002 Jun [en línia] http://genome.ucsc.edu/cgi-bin/hgGateway

[25] BLAT: Kent WJ. BLAT - the BLAST-like alignment tool. *Genome Res.* 2002 Apr [en línia] http://genome.ucsc.edu/cgi-bin/hgBlat

[26] HUGO *Gene Nomenclature Committee at the European Bioinformatics Institute* [en línia] http://www.genenames.org/

[27] COSMIC *Catalogue Of Somatic Mutations In Cancer* [en línia] http://cancer.sanger.ac.uk/cosmic/